Master mathématiques appliquées, statistiques

Parcours statistiques et data science, ingénierie mathématique

Deuxième année

# Proposition d'une approche pour vérifier la cohérence entre données opportunistes et données protocolées pour mieux estimer l'état et la dynamique de la biodiversité

---

# Rapport de stage

---

*Etudiant* : Pierre Bouchet bouchetpierre15@gmail.com

*Tuteur universitaire* : Didier Chauveau didier.chauveau@univ-orleans.fr

*Maître de stage* : Frédéric Gosselin frederic.gosselin@inrae.fr

INRAE UR EFNO – Nogent-sur-vernisson, France

2 Mars 2020 – 28 Août 2020

*« All models are wrong, but some are useful »*

***George Box***

*« Oui mais, quand je reste trop dans ma baraque, je conspire, c'est un réflexe. Du coup, je prends l'air, ça vaut mieux pour tout le monde.* »

***François Rollin dans Kaamelott, écrit par Alexandre Astier***

# Table des matières

## Remerciements

Mon entière reconnaissance va tout naturellement à Frédéric Gosselin pour m'avoir proposé un sujet d'étude aussi riche. Je tiens à le remercier pour la confiance qu'il m'a accordée et pour les discussions que nous avons eues. J'admire son savoir et lui suis reconnaissant de m'avoir permis de faire mes premiers pas en recherche avec lui et de surcroît de m'avoir fait découvrir la recherche en écologie, domaine on ne peut plus actuel.

Je souhaiterais aussi remercier l'ensemble du personnel de l'INRAE pour m'avoir accueilli dans une ambiance amicale et détendue, bien que la crise sanitaire ait réussi à confiner tout le monde chez soi durant une majeure partie de mon stage, raccourcissant ma présence à Nogent-sur-Vernisson. Je pense surtout à Sylvie avec qui j'ai pu discuter de nombreuses fois, à Guilhem pour sa sympathie et pour son invitation au badminton et enfin plus largement tous les présents au badminton le mardi et jeudi midi : merci.

Merci également à Didier Chauveau pour son aide, et à Mame Diarra Fall pour son cours de statistiques bayésiennes et de statistiques en troisième année de licence qui m'a fait découvrir un domaine qui m'était étranger auparavant, et sans laquelle je ne me serais (presque sûrement) jamais destiné aux statistiques.

# 1. Introduction

## 1.1. INRAE

Du 2 Mars 2020 au 28 Août 2020 j'ai pu réaliser mon stage de fin d'études de Master au sein de l'INRAE sous la direction de Frédéric Gosselin, ingénieur-chercheur en écologie forestière et biométrie.

L'Institut National de Recherche pour l'Agriculture, l'alimentation et l'Environnement est un organisme de recherche français, comptant parmi les premiers instituts au monde pour l'étude des relations agriculture, environnement et alimentation et est sous la double tutelle du ministre chargé de la Recherche et du ministre chargé de l'Agriculture.

J'ai intégré le centre de Nogent-sur-Vernisson sur le domaine des Barres. Les recherches portent sur les écosystèmes forestiers de plaine et les pratiques de gestion sylvicole favorables à la préservation de la biodiversité forestière. Les études réalisées conduisent à des modèles théoriques et des publications scientifiques, mais l'institut est également tourné vers l'appui aux décideurs publics. Plusieurs équipes de recherche sont présentes dont l'équipe "Biodiversité" à laquelle j'ai été intégré. L'équipe centre ses activités de recherche sur des composantes de la biodiversité dans les massifs forestiers afin de mieux comprendre la biodiversité ainsi que les pressions qui pèsent dessus dans le but de proposer des recommandations de gestion forestière et d'aménagement du territoire préservant la biodiversité.
L'un des projets menés par l'institut est nommé PASSIFOR 2 et c'est dans ce projet que s'insère mon stage.

## 1.2. Le projet PASSIFOR2

Le projet PASSIFOR2 (Proposition d'Amélioration du Système de Suivi de la biodiversité FORestière), financé par le ministère de la Transition Ecologique et Solidaire, constitue la phase 2 du projet PASSIFOR (2011-2015) soutenu par le ministère de l'agriculture.

L'objet de ce projet est d'élaborer différents assemblages d'éléments existants et à créer – appelés ici « maquettes » – de suivi de la biodiversité en forêt. Il vise une aide aux politiques publiques dans le domaine du suivi continu de la biodiversité, centré sur la forêt en lien avec les autres milieux. En dépit d'acquis importants dans le domaine, les indicateurs actuels de biodiversité forestière sont surtout des indicateurs indirects, ciblés sur les habitats d'espèces et mobilisant principalement des données dendrométriques (mesures caractéristiques physiques quantifiables des arbres) ; il importe d'acquérir des informations qui permettent de (i) mieux cerner directement l'état et la dynamique de la biodiversité forestière et (ii) mieux évaluer le lien entre politiques publiques en forêt, pratiques de gestion et biodiversité.

Ce projet est constitué de cinq tâches et le stage fait partie de la tâche E nommée *mesures, échantillonnage, analyses statistiques*. Cette dernière a pour objectif de fournir les éléments relatifs à **l'échantillonnage** (où faire les relevés de biodiversité), à la **mesure** (comment relever la biodiversité) et à **l'analyse des données** de biodiversité.

Ce projet est mené conjointement entre l'INRAE, l'UMS PATRINAT, le GIP ECOFOR et les responsables sont Frédéric Gosselin, ingénieur chercheur en écologie forestière ainsi que Guy Landmann qui est directeur-adjoint du GIP ECOFOR.

L'UMS PATRINAT (Unité Mixte de Service Patrimoine Naturel) assure des missions nationales d'expertise scientifique et de gestion des connaissances en biodiversité pour ses trois tutelles, que sont l'Office français de la biodiversité (OFB), le Muséum national d'Histoire naturelle (MNHN), et le Centre national de la recherche scientifique (CNRS).

Le GIP ECOFOR (Groupement d'Intérêt Public Ecosystèmes forestiers) est un groupement d'intérêt public comptant actuellement comme membres les principaux organismes publics français actifs dans le domaine de la recherche forestière (AgroParisTech, INRAE, CNRS, IGN, OFB, ONF, MNHN, …) et l'État représenté par les deux ministères chargés de l'agriculture et de la forêt d'une part et celui de la transition écologique et solidaire d'autre part. Ses activités principales portent sur l'étude du fonctionnement et de la dynamique des écosystèmes, ainsi que sur la gestion durable des forêts, le tout en milieux tempérés et tropicaux. Le GIP Ecofor joue le rôle d'interface entre recherche et gestion forestière, grâce à son expertise et au moyen de systèmes d'informations spécialisés.

## 1.3. Problématique

La prise en compte et l'évaluation de la biodiversité sont devenus des enjeux pour nos sociétés suite au sommet de la Terre à Rio en 1992. Par ailleurs la biodiversité a été définie à cette occasion comme «la variabilité des organismes vivants de toute origine y compris, entre autres, les écosystèmes terrestres, marins et les autres écosystèmes aquatiques et les complexes écologiques dont ils font partie; cela comprend la diversité au sein des espèces, et entre espèces ainsi que celle des écosystèmes» (https://www.cbd.int/doc/legal/cbd-fr.pdf). Cependant le suivi de la biodiversité est beaucoup moins développé que le suivi du climat. Non seulement parce que les suivis de biodiversité ont commencé plus tard mais aussi car il s'agit d'une entité très multivariée. On se confronte aujourd'hui à un problème de trous dans les suivis et les indicateurs de biodiversité.
La biodiversité est étudiée à travers plusieurs aspects :

- L'abondance, notée $N_i$ pour l'espèce $i$ et qui peut prendre des formes différentes, non «équivalentes les unes aux autres : (i) l'abondance numérique, i.e, le nombre d'individus de l'espèce (ii) la biomasse de l'espèce (iii) le recouvrement en pourcentage de la surface recouverte par l'espèce
- La richesse qui correspond au nombre d'espèces différentes correspondant à une groupe taxonomique donné et à un lieu fixé. Il s'agit également de données de comptage.
- La présence ou présence/absence qui indique si une espèce donnée est présente uniquement ou dans le cas d'une présence/absence on étudie si l'espèce est présente ou absente. Il s'agit dans ce cas de données binaires 0 ou 1.

Il existe évidemment d'autres aspects suivant l'étude que l'on souhaite mener (comme les aires de répartitions), mais nous allons nous intéresser à l'abondance relative, définie ainsi : $N_{ij}/N_{i1}$ pour tout $i$ et pour tout $j$ où $N_{ij}$ représente le nombre d'individus de l'espèce $i$ dans le site $j$.

Nous l'avons dit, un des problèmes avec les données de biodiversité est que les données sont rares. Plus précisément, les données ayant suivi un protocole d'échantillonnage sont rares. Une des façons d'augmenter le nombre de données est de considérer les données de sciences participatives, ou sciences citoyennes. Un exemple de données citoyennes est le *Breeding Bird Survey* où les observations étaient faîtes par des volontaires donnant le nombre d'individus de chaque espèce d'intérêt observés à un endroit donné et dans un intervalle de temps donné (Johnson, 2007; Link & Sauer, 1998).

Le problème avec ce type de données citoyennes ou participatives est qu'elles ont des propriétés statistiques non contrôlées :

- Biais géographique : lieux non représentatifs ou déjà visités (van Strien, van Swaay, & Termaat, 2013).
- Biais d'observation : intensité de recherche inconnue (van Strien et al., 2013)
- Biais de report : non signalement d'une espèce jugée inintéressante par l'observateur (van Strien et al., 2013)
- Biais d'échantillonnage : les personnes vont aller sur des endroits plus accessibles pour eux (bord d'une route etc.) (van Strien et al., 2013). Le biais géographique et le biais d'échantillonnage peuvent se ressembler mais diffèrent : le premier porte sur les coordonnées x,y et le second porte sur le type de milieu. On pourrait toutefois les regrouper en un seul biais appelé « biais spatial ».

Bien que parfois ces données, que nous appellerons désormais « opportunistes » soient une source potentiellement fiable d'information de changements sur les distributions d'espèces (Schmeller et al., 2009), il arrive aussi que ces données causent des tendances fallacieuses ou masquent des tendances réelles (Dennis & Thomas, 2000; Kéry et al., 2010; Kuussaari, Heliölä, Pöyry, & Saarinen, 2007; Szabo, Vesk, Baxter, & Possingham, 2010).
Par exemple, (Snäll, Kindvall, Nilsson, & Pärt, 2011) a démontré une faible correspondance entre les tendances des données opportunistes et les tendances de données standardisées d'oiseaux en Suède.

Le but global est de mettre en place un outil pour détecter si les données opportunistes peuvent être adjointes aux données protocolées. De nombreux articles ont proposé des méthodes pour coupler données opportunistes et données protocolées. Dans ce stage nous allons nous intéresser aux travaux de (Giraud, Calenge, Coron, & Julliard, 2015) et de (Coron, Calenge, Giraud, & Julliard, 2018). Plus particulièrement, nous allons utiliser une variante de ces méthodes en y ajoutant des *goodness-of-fit* p-values (Gosselin, 2011; Johnson, 2007; Robins, van der Vaart, & Ventura, 2000). De plus nous utiliserons une approche bayésienne pour l'estimation, nous aurons donc recours à des algorithmes MCMC. Toutes les analyses seront faites avec le logiciel R 3.6.3 (Team, 2020) et en particulier pour la partie MCMC nous utiliserons JAGS (Martyn Plummer, 2004) ainsi que plusieurs packages R, à savoir rjags (M. Plummer, 2019), coda (M. Plummer, Best, Cowles, & Vines, 2006) et rstan (Stan Development, 2020).

# 2. Matériels et méthodes

## 2.1. Modèle statistique de Coron et al 2018

Comme nous l'avons dit dans l'introduction, nous allons utiliser pour l'estimation statistique la méthode proposée par (Coron et al., 2018) et qui fait suite à la publication de (Giraud et al., 2015). Nous allons donc synthétiser les idées de ce modèle dans cette section.

Le but de l'article est de développer une procédure statistique pour estimer l'abondance relative d'espèces ainsi que leurs préférences pour habitats respectives, tout cela en utilisant des données opportunistes. Ici on utilise trois jeux de données : un protocolé, un opportuniste et un de test. Nous l'avons vu, les données opportunistes comportent des biais inhérents et le but d'un tel modèle est donc de les corriger grâce aux données protocolées. Nous allons d'abord décrire les ingrédients écologiques, qui ne sont pas dépendants du type du jeu de données, et ensuite les ingrédients d'observation qui eux sont dépendants du type de jeu de données.

### 2.1.1. Abondances et probabilité de sélection d'habitat

L'espace est divisé en unités appelées *sites* et sont indexés par $j \in [\![1; J]\!]$ et les espèces sont indexées par $i \in [\![1; I]\!]$ . On note $N_{ij}$ le nombre d'individus de l'espèce $i$ sur le site $j$ et l'objectif est d'estimer les abondances relatives $N_{ij}/N_{i1}$ pour tout $i$ et pour tout $j$. Le choix du site 1 comme référence est totalement arbitraire.

Le type d'habitat d'un site $j$ n'est pas homogène : chaque site est composé de domaines qui représentent chacun un habitat. On note $h \in [\![1; H]\!]$ l'habitat. De plus les espèces ne sont pas uniformément distribuées dans les domaines : chaque espèce a un habitat de prédilection et donc sera plus ou moins présente suivant l'habitat. C'est pourquoi on introduit une probabilité de sélection d'habitat $h$ par l'espèce $i$ notée $S_{ih} \in [0,1]$.

Ainsi le modèle suppose que la densité d'animaux de l'espèce $i$ présents dans la localisation $x$ au sein du site $j$ est donnée par :

$$\frac{N_{ij} S_{ih(x)}}{\sum_{h'} S_{ih'} V_{h'j}}$$

où $h(x)$ est le type d'habitat à la localisation $x$, $V_{hj}$ est l'aire de l'habitat de type $h$ au sein du site $j$ et $S_{ih} \in [0,1]$ est la probabilité de sélection de l'habitat de type $h$ par l'espèce $i$. Notons que ces probabilités de sélection d'habitat sont inconnues et seront estimées.

### 2.1.2. Observations et rapport

Comme nous l'avons déjà dit, l'estimation repose sur deux jeux de données : l'un ayant suivi un protocole strict labellisé $k = 0$ et l'autre étant caractérisé par un effort d'échantillonnage inconnu, labellisé $k = 1$. On suppose que chaque site $j$ a été échantillonné par les deux jeux de données et que chaque espèce $i$ a été échantillonnée dans au moins l'un des deux jeux de données.

Ensuite on divise chaque site $j$ en cellules notées $c$, et pour chaque observation, on connaît la cellule correspondante. Il est important de noter que le pavage de cellules peut totalement différer entre les

deux jeux de données et que dans chaque jeu de données seule une (possiblement faible) proportion de cellules a été visité par au moins un observateur. Notons $X_{ick}$ le comptage de l'espèce $i$ dans la cellule $c$ pour le jeu de données $k$. Ce comptage est biaisé par l'inhomogène effort d'observation (temps d'observation, nombre d'observateurs, nombre de pièges, etc.) et par la probabilité de rapport (détectabilité variable, etc.).

Notons $E_{ck} \in \mathbb{R}$ l'intensité d'effort dans la cellule $c$ pour le jeu de données $k$ et $P_{ik} \in \mathbb{R}$ la probabilité de détection de l'espèce $i$ pour le jeu de données $k$ avec la convention que si l'espèce $i$ n'a pas été surveillée dans le jeu de donnée $k$ alors on fixe $P_{ik} = 0$. On suppose de plus que les observateurs ne parcourent pas l'espace uniformément : ils ont des préférences pour certains types d'habitats et qui ne sont pas les mêmes suivant le type du jeu de données. Ces préférences induisent alors un biais et c'est pourquoi, de la même façon que nous avons introduit une probabilité de sélection d'habitat, nous allons introduire une préférence des observateurs pour l'habitat $h$ du jeu de données $k$ notée $q_{hk} \in [0,1]$. Ainsi pour le jeu de données $k$ on modélise l'intensité d'observation de la location $x$ au sein de la cellule $c$ par :

$$\frac{q_{h(x)k} E_{ck}}{\sum_{h'} q_{h'k} V_{h'c}}$$

où $V_{hc}$ est l'aire connue de la cellule $c$ couverte de l'habitat $h$. Notons que dans notre modèle on suppose que les observateurs ont tous la même préférence par habitat : il y a homogénéité entre observateurs des préférences d'habitat (par exemple ils vont tous préférer les forêts de conifères).

## 2.2. Modèle statistique

Après avoir posé les ingrédients écologiques et ensuite les ingrédients d'observation, nous pouvons expliciter le modèle final :

$$X_{ick} \sim \mathcal{P}oisson\left( N_{ij} E_{ck} P_{ik} \sum_{h} \frac{q_{hk}}{\sum_{h'} q_{h'k} V_{h'c}} \times \frac{S_{ih}}{\sum_{h'} S_{ih'} V_{h'j}} V_{hc} \right)$$

On rappelle que dans l'équation ci-dessus les volumes $V_{hj}$ et $V_{hc}$ sont connus. De plus, pour le jeu de données standardisé, l'intensité d'observation $E_{c0}$ est connue et qu'on suppose que (i) soit le type d'habitat associé à chaque observation $X_{ic0}$ est connu (dit autrement que les cellules du jeu de données standardisé sont suffisamment petites pour n'être constituées que d'un seul habitat), (ii) soit le ratio $q_{h0}/q_{10}$ est connu pour tout $h$ (et vaut généralement 1). Tous les autres paramètres sont inconnus.

Ainsi nous devons estimer cinq matrices de paramètres, mais par souci d'identifiabilité du modèle statistique, certains paramètres seront fixés :

- $N_{ij}, 1 \leq i \leq I,\ 1 \leq j \leq J$ qui correspond à l'abondance réelle sous-jacente du site et qui est non observée. Cette matrice est entièrement estimée.
- $E_{ck}, 1 \leq c \leq 40J,\ 1 \leq k \leq K$, l'effort. La première colonne ainsi que les $10 \times J$ premières lignes de la deuxième colonne sont fixés.

- $P_{ik}, 1 \leq i \leq I,\ 1 \leq k \leq K$ correspondant à la probabilité de reporting d'espèce par l'observateur. La première colonne ainsi que le premier élément de la deuxième colonne sont fixés.
- $S_{ih}, 1 \leq i \leq I,\ 1 \leq h \leq H$ paramètre de sélection d'habitat par les espèces. La première colonne de cette matrice est fixée.
- $q_{hk}, 1 \leq h \leq H,\ 1 \leq k \leq K$ qui correspond à la préférence d'habitat pour les observateurs. La première colonne ainsi que le premier élément de la deuxième colonne sont fixés.

De façon générale, sont fixés les paramètres correspondant au jeu de données protocolées car dans un contexte plus appliqué, c'est ce qui est fait.

### 2.3. Hypothèses du code R de simulation

Nous venons d'aborder le modèle statistique utilisé, nous allons désormais évoquer les hypothèses faites dans le code de simulation R des données utilisées dans la phase de simulation du travail de Coron et al. (2018) dont nous utiliserons une version légèrement modifiée. En effet, certaines hypothèses peuvent être fortes d'un point de vue écologique et serviraient alors par la suite de base de réflexion pour des scenarii de simulation.

Dans le code nous posons $I = 20, J = 30$ et $K = 2$. Nous étudions donc 20 espèces sur 30 sites avec 2 jeux de données. La première hypothèse est que chaque site $j \in \{1, 2, \dots, J\}$ est découpé en 40 cellules réparties comme suit : 10 cellules de données protocolées et 30 de données opportunistes. Ceci aboutit à un découpage uniforme de tous les sites. Or, dans un contexte appliqué, il se pourrait que (i) certains sites contiennent plus de relevés que d'autres (cas de sites situés aux frontières pour les données standardisées et variation de l'effort opportuniste entre sites), que (ii) ce soit une autre proportion que celle 1 pour 3 entre les données protocolées et opportunistes qui prévale ou (iii) que cette même proportion soit variable d'un site à l'autre.

Ensuite nous posons $H = 2$ ce qui indique que nous ne considérons que 2 habitats ce qui est peu nuancé et correspondrait par exemple dans un cas forestier à forêts feuillues vs forêts de résineux. De plus on suppose que les habitats sont tirés aléatoirement dans une même loi uniforme ne dépendant pas du site $j$.

Puis pour chaque cellule (au nombre de $40 \times J = 1200$ ) l'aire de chaque habitat est fixée à 1km$^2$ pour les données protocolées, correspondant à un effort contrôlé et équivalent pour les différents habitats, et sont aléatoires pour les données opportunistes.

Une autre hypothèse forte est que chaque site standardisé, d'aire 1km$^2$, est dans un habitat pur. Or dans un contexte plus appliqué on serait plus probablement avec un mélange d'habitats dans une surface de cette échelle.

La sélection d'habitat par les espèces fait l'hypothèse d'être constante entre les sites j pour chaque espèce, ce qui ne correspond pas à toutes les situations écologiques, où certaines espèces peuvent avoir des préférences d'habitat qui varient dans l'espace.

On peut évoquer l'hypothèse faite que les préférences des observateurs pour les habitats sont les mêmes pour tous les observateurs et sont fixées à 1 pour les données protocolées. Or tous les observateurs ne peuvent pas avoir exactement les mêmes préférences d'habitats les uns par rapport aux autres. En particulier dans un panel d'observateurs « amateurs » (typiquement des citoyens bénévoles) il est très peu probable de rassembler que des personnes avec les mêmes préférences d'habitat : il y aura forcément des différences.

De plus nous faisons l'hypothèse que l'abondance en espèces est uniforme sur les sites et sur les espèces ce qui constitue une hypothèse très forte. Cela signifie par exemple qu'il n'y a pas – en moyenne – d'espèce plus abondante que les autres – et donc pas d'espèce plus rare. En outre, comme introduit dans le modèle statistique plus haut, l'abondance sous-jacente est modélisée par

$N_{ij}$ et dépend donc uniquement de l'espèce $i$ et du site $j$, mais ne dépend pas de la cellule $c$. Dit autrement, l'abondance sous-jacente a une variabilité **inter**-sites mais pas **intra**-sites. C'est-à-dire qu'au sein d'un site, l'abondance sous-jacente est la même pour toutes les cellules, ce qui constitue une hypothèse forte écologiquement parlant. Par ailleurs, le tirage au sort de l'abondance locale suit une loi Poissonnienne, alors que certaines espèces vivent par groupes et pourraient avoir des abondances par « grappes » non uniformes tandis que d'autres espèces sont territoriales et pourraient être réparties plus régulièrement que supposé par la loi de Poisson.

Il y a également les hypothèses sur la probabilité de rapport (ou l'anglophile « reporting ») et de détection : on suppose qu'elle est issue d'une loi uniforme sur les espèces et sur les jeux de données. De plus elle ne dépend ni de l'observateur (et de sa qualité par extension) ni de l'habitat, or il se pourrait qu'un habitat défavorise ou favorise la probabilité de détection d'une espèce. Aussi la probabilité de rapport est la même pour les deux jeux de données et on pourrait – dans un contexte appliqué – s'attendre à ce que les probabilités de rapport soient plus faibles pour les données opportunistes que pour les données protocolées.

Pour finir l'effort pour le jeu de données protocolées est posé comme valant 0 sur les cellules des données opportunistes et vaut 1 (donc faible et standardisé) pour les cellules des données protocolées. De plus l'effort pour le jeu de données opportunistes est posé comme valant 0 sur les cellules des données protocolées, est uniformément réparti et très élevé (en moyenne 50) pour les cellules des données opportunistes. Cela signifie qu'il y a un rapport de 50 pour 1 – en moyenne – entre effort des données opportunistes et effort des données protocolées. De surcroît l'effort suit la même loi uniforme quel que soit le site $j$.

Pour conclure cette section, nous avons exposé le modèle statistique et ses hypothèses. Pour résumer les résultats obtenus par (Coron et al., 2018), on observe une meilleure estimation (moins de biais et de variance) ainsi que de meilleures performances prédictives en fusionnant les données protocolées et opportunistes plutôt qu'en utilisant uniquement les données protocolées. Avant d'aborder la méthodologie utilisée pour ce stage, nous allons poser le cadre statistique englobant : les statistiques bayésiennes.

## 2.4. Inférence bayésienne

### 2.4.1. Généralités sur l'inférence bayésienne

On rappelle que nous sommes dans un contexte bayésien et dans cette section nous allons poser des généralités ainsi que les notations afin que le lecteur puisse avoir une base pour poursuivre. Nous invitons le lecteur, pour une lecture plus approfondie et exhaustive, l'ouvrage suivant (Gelman et al., 2014).

Soit $y_{obs} \in \mathbb{R}^n$ les données observées, $\theta \in \Theta \subset \mathbb{R}^p$ le vecteur de paramètres (incluant les hyperparamètres si l'on utilise un modèle hiérarchique) que l'on cherche à estimer. L'analyse statistique bayésienne a pour but d'exploiter de la façon la plus efficace possible l'information apportée par les données $y_{obs}$ sur le paramètre $\theta$ afin de construire des procédures d'inférence dessus. L'information fournie par les données est contenue dans la vraisemblance notée $g(y_{obs}|\theta)$.

Dans le paradigme bayésien, le paramètre $\theta$ inconnu est vu comme une variable aléatoire. L'espace des paramètres $\Theta$ est alors muni d'une loi *a priori* notée $\pi(\theta)$. Cela revient à supposer que $\theta$ est distribué suivant $\pi$ « avant » que les données ne soient générées suivant $g(.|\theta)$. Le choix de la loi *a priori* est une étape fondamentale. On entend par information *a priori* sur $\theta$ toute information en dehors de celle apportée par les observations. Ainsi le choix de la loi *a priori* devrait être basé sur les expériences du passé, un avis d'expert, une intuition etc. Ce choix est toutefois assez difficile à réaliser en pratique car d'une part il peut être compliqué de résumer toute l'information disponible sur $\theta$ au travers d'une seule distribution. Et d'autre part le choix de la loi *a priori* influe sur le reste : des lois *a priori* différentes peuvent mener à des conclusions divergentes ! C'est d'ailleurs la principale critique faite aux statistiques bayésiennes. En pratique le choix de la loi *a priori* peut être motivé par des aspects de calculabilité avec les lois conjuguées par exemple. Ou pour tenir compte du fait que l'on ne sait pas grand-chose sur le paramètre, on opte pour des lois à faible contenu informatif. Il existe par ailleurs la règle de Jeffreys pour déterminer des lois *a priori* dont le principe est de favoriser l'information apportée par les données.

L'inférence sur le paramètre $\theta$ repose sur la loi *a posteriori*. Cette dernière joue un rôle primordial et peut s'interpréter comme une mise à jour de la loi *a priori* une fois que les données ont été observées. La densité de la loi *a posteriori* est donnée, grâce à la formule de Bayes (d'où le qualificatif) ainsi :

$$\pi(\theta|y_{obs}) = \frac{g(y_{obs}|\theta)\pi(\theta)}{\int_\Theta g(y_{obs}|\theta)\pi(\theta)d\theta}$$

Le dénominateur de la formule de la loi *a posteriori* ne dépend pas de $\theta$. Or la loi *a posteriori* est une fonction de $\theta$. Ainsi le dénominateur est par rapport à $\theta$ une constante et on peut donc réécrire la loi *a posteriori* avec la notation proportionnelle :

$$\pi(\theta|y_{obs}) \propto g(y_{obs}|\theta)\pi(\theta)$$

La loi *a posteriori* permet de calculer un estimateur bayésien de $\theta$ suivant la fonction de coût que l'on applique. Sous l'hypothèse d'un coût quadratique, l'estimateur bayésien de $\theta$ est la moyenne de la densité *a posteriori* $\mathbb{E}^{(\theta|y_{obs})}[\theta] = \int_\Theta \theta d\pi(\theta|y_{obs})$. Il est d'ailleurs possible de construire le pendant bayésien des tests d'hypothèses ou intervalles de confiance fréquentistes.

Toutefois, l'expression de la loi *a posteriori* est souvent compliquée à obtenir explicitement et on a alors recours à des algorithmes MCMC (Monte Carlo Markov Chain). Le principe d'un MCMC est

d'obtenir un échantillon d'une loi cible à l'aide de chaînes de Markov dont la distribution stationnaire est la loi cible que l'on cherche à échantillonner. Les méthodes les plus connues d'échantillonnage MCMC sont Métropolis-Hastings (Chib & Greenberg, 1995) et l'échantillonneur de Gibbs (Casella & George, 1992). Il faut donc s'assurer que les chaînes de Markov (il peut n'y en avoir qu'une) aient bien convergées vers la distribution stationnaire !

### 2.4.2. Etude de la convergence des MCMC

Les vérifications de la convergence d'un MCMC sont assez heuristiques et pas entièrement rigoureuses mais elles permettent de s'assurer une certaine véracité des valeurs.

La première étape est ainsi la vérification de la convergence. Ceci se fait de plusieurs façons : on peut tracer les trajectoires de chaque chaîne et vérifier que les chaînes se ressemblent après une période que l'on appelle en anglais « *burn-in* » et qui correspond à la période transitoire de la chaîne de Markov. En traçant les trajectoires de chaque chaîne on a autant de graphiques à regarder que de paramètres estimés ! Ce qui peut vite s'avérer fastidieux si le nombre de paramètres est de l'ordre de la centaine.

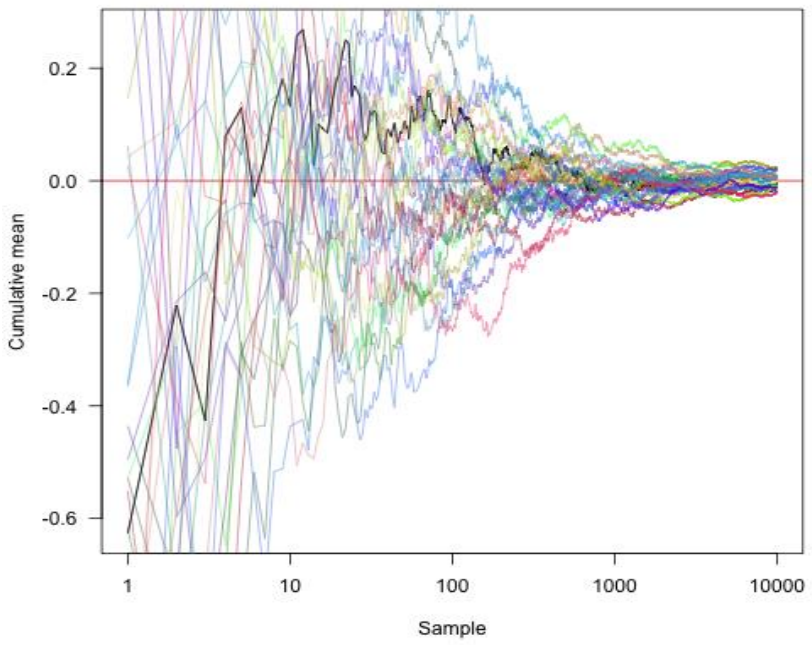


*Figure 1 - Exemple de trajectoire de MCMC*

Par exemple dans la Figure 1 - Exemple de trajectoire de MCMC on peut voir plusieurs chaînes de Markov et graphiquement le burn-in est d'environ 5000. Graphiquement cela correspond au moment où la trajectoire se stabilise. Il est important de noter que dans cet exemple il y a beaucoup de chaînes de Markov mais qui sont toutes utilisées pour l'échantillonnage d'un seul paramètre.

On peut alors utiliser un autre outil qui est numérique : le diagnostic de convergence de Geweke. Nous utilisons ce diagnostic car nous n'allons utiliser qu'une seule chaîne de Markov et que les autres outils sont utilisables dans le cas de plusieurs chaînes de Markov. Proposé en 1991 par Geweke (Geweke, 1991), ce diagnostic consiste à séparer en deux groupes les valeurs de chaque chaîne :

un premier groupe est formé par les 10% premières itérations et le second groupe est constitué des 50% dernières itérations, passant à la trappe 40% des itérations. Ensuite l'idée est la suivante :

Si nous sommes à l'état stationnaire de la chaîne, alors les moyennes des deux groupes devraient être au moins semblables. On fait alors un test de Welch dont la statistique est la suivante :

$T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{[\frac{s_1^2}{n_1} - \frac{s_2^2}{n_2}]}}$ avec $\overline{X_i}, s_i^2, n_i$ représentant respectivement la moyenne, la variance et la taille de l'échantillon $i$ pour $i \in \{1,2\}$. Cette statistique de test suit sous l'hypothèse nulle d'égalité des moyennes une loi normale centrée-réduite. La variance est estimée via une densité spectrale ce qui permet de prendre en compte l'autocorrélation inhérente aux MCMC.

Encore une fois, si nous avons beaucoup de paramètres à estimer, admettons 1000, alors vérifier 1000 réalisations de la statistique $T$ serait trop fastidieux. La vérification consiste à étudier si oui ou non ces 1000 valeurs sont « cohérentes » avec une loi normale centrée-réduite en la comparant au quantile d'ordre $1 - \alpha$ où $\alpha$ représente le risque de première espèce. Donc pour être plus efficace nous allons regarder la proportion de valeurs de statistique qui sont dans l'intervalle $I = [-1.96, 1.96]$ qui correspond à l'intervalle formé par le quantile d'ordre 0.025 et d'ordre 0.975 d'une loi normale centrée-réduite. Ainsi, si nous avons une proportion d'environ 95% de valeurs qui se situent dans l'intervalle $I$ alors nous considérerons que l'hypothèse nulle est vérifiée, ce qui indique que les moyennes des deux groupes sont égales et donc que les deux groupes sont issus de la phase stationnaire de la chaîne. Au contraire, si l'on observe une proportion significativement inférieure à 95% alors cela indique que les 10% premières itérations sont issues de la phase de burn-in. Cela voudra donc dire que l'on doit ajouter à la phase de burn-in les 10% premières itérations. Une fois que la phase de burn-in est passée, la chaîne de Markov atteint ainsi sa distribution stationnaire.

Une fois que la convergence est vérifiée, à travers le burn-in, on passe à la 2e vérification : l'indépendance. En effet il faut s'assurer que les valeurs des paramètres peuvent être considérés comme des tirages indépendants dans la distribution postérieure car sinon c'est de la pseudo-réplication. Attention : on ne passe à l'indépendance qu'une fois que la convergence est vérifiée et correcte ! Il est inutile de vérifier l'indépendance sur une chaîne qui n'est pas convergente. La principale source d'autocorrélation est la dépendance entre les itérations de la chaîne de Markov.

Pour étudier la question de l'autocorrélation il y a 3 outils principaux :

i. La comparaison de deux erreurs-types : « naive SE » et « time-series SE ».

   La première est définie ainsi : $SE_{naive} = \sqrt{\frac{Var(X)}{C*S}}$ avec $C$ le nombre de chaînes, $S$ le nombre d'itérations par chaîne et $X$ la concaténation vectorielle des valeurs de chaque chaîne.

   La deuxième erreur-type est définie par : $SE_{time-series} = \sqrt{\frac{Var_{ts}(X)}{C*S}}$ où $Var_{ts}(X)$ est la moyenne de $Var_{ts}\left(X^{(c)}\right)^{(c)}$ pour chaque ensemble d'échantillons $X^{(c)}$ de chaque chaîne et chaque $Var_{ts}\left(X^{(c)}\right)^{(c)}$ est obtenu en appliquant un modèle AutoRégressif sur $X^{(c)}$ dont l'ordre est choisi via l'AIC (*Akaike Inforation Criterion*).

   Ainsi, plus le terme $SE_{time-series}$ est proche de $SE_{naive}$ moins il y a d'autocorrélation.

ii. Le nombre de valeurs efficaces. Il est défini par : $N_{eff} = \frac{n}{1+2\sum_{k=1}^{\infty}\rho(k)}$ avec $n$ le nombre d'itérations par chaîne et $\rho(k)$ l'autocorrélation au décalage d'ordre $k$.
L'idée est d'avoir une sorte de « taux de change » entre échantillon indépendant et dépendant. Par exemple on pourrait vouloir que nos 1000 itérations MCMC « valent » 80 échantillons indépendants de la distribution postérieure, car rappelons-le, les échantillons MCMC sont corrélés.
Attardons-nous maintenant sur la formule. Si les échantillons sont indépendants, alors $N_{eff} = n$. Si la corrélation au décalage $k$ décroît très lentement en fonction de $k$, alors la somme diverge et on a $N_{eff} \approx 0$.

Ainsi toute chaîne de Markov raisonnable est entre ces deux extrêmes. Bien sûr on veut éviter l'autocorrélation entre les échantillons du MCMC mais il s'agit d'un vœu pieu : on ne pourra pas avoir une autocorrélation nulle. Cependant l'idéal est d'avoir une autocorrélation qui vaut rapidement 0 de sorte que la série numérique au dénominateur de l'équation de $N_{eff}$ soit convergente mais convergente vers une valeur « pas trop grande » (Kass, Carlin, Gelman, & Neal, 1998).

iii. Le graphique d'autocorrélation. Ce graphique trace l'autocorrélation pour chaque paramètre et avec une couleur par chaîne de Markov. Le but est de repérer pour quel décalage (ou « lag » en anglais) la valeur d'autocorrélation est minimale en valeur absolue.

Avec ces trois outils on peut trouver ce que l'on appelle le « thin » qui correspond à la période avec laquelle on garde les itérations du MCMC. Par exemple, prenons un MCMC de 10 000 itérations, avec un burn-in de 1000 et un thin de 10. Cela veut dire que sur 10 000 itérations de départ, nous en enlevons 1000 qui correspond au burn-in et ensuite sur les 10 000 – 1000 itérations qu'il nous reste nous en prendrons qu'1 sur 10. Ainsi au final nous aurions 900 itérations.

Une fois que le modèle statistique a été ajusté on peut se poser une question à savoir : « Les données observées sont-elles compatibles avec le modèle ? ». C'est à cette question que l'on essaie de répondre avec les goodness-of-fit juste après.

## 2.5. Goodness of fit p-values

### 2.5.1. Introduction

La critique de modèle statistique a été déclarée par George Box comme l'une des deux principales étapes lors du développement de modèle statistique (Box, 1980) et a d'ailleurs créé un bien célèbre aphorisme : « *All models are wrong, but some are useful* » (Box, 1976).

De nombreux termes sont utilisés pour désigner la comparaison entre données observées et modèle ajusté – adéquation du modèle, vérification du modèle, validation du modèle, évaluation du modèle (Gelman et al., 2014; O'Hagan, 2003) – mais nous utiliserons dorénavant le terme « qualité d'ajustement » ou plus précisément le terme anglais « goodness-of-fit » abrégé en **GOF**. A date il existe deux types de GOF : externe ou interne. En ce qui concerne la première, le principe est de séparer les données en deux (schéma que l'on retrouve en machine learning avec le découpage apprentissage-test) et d'utiliser une partie des données pour ajuster le modèle et l'autre partie des données (inutilisées jusqu'alors) est utilisée en confrontation avec le modèle avec comme question de principe : « Est-ce que les données sont cohérentes avec le modèle ? ». Le terme « cohérent » est quantifié à l'aide d'une fonction de divergence (« *discrepancy function* » en anglais) qui dépend des paramètres du modèle statistique et des données. On peut remarquer qu'un cas particulier de fonction de divergence qui ne dépend que des données est une statistique. L'autre type de GOF, interne, utilise toutes les données pour ajuster le modèle puis réutilise les données pour les confronter aux prédictions du modèle : on utilise les données deux fois. Pour notre part nous ne nous intéresserons qu'aux GOF internes qui présentent un clair avantage de permettre l'utilisation complète des données pour la construction du modèle. Plus précisément nous allons nous concentrer sur les GOF p-values qui est un outil numérique et univarié. Ces GOF p-values sont « Fishériennes », c'est-à-dire qu'elles représentent « la probabilité de voir quelque chose au moins aussi étrange que ce que l'on voit déjà » (Christensen, 2005). Ainsi les GOF p-values ont de Fishérien qu'elles comparent le modèle aux données, contrairement à l'approche de Neyman-Person qui compare deux modèles ou deux hypothèses mutuellement incompatibles.

Pour résumer, les GOF ont pour but de juger de la qualité absolue du modèle ajusté. Nous allons dans les sections suivantes, détailler le fonctionnement de deux GOF p-values dont celle que nous allons utiliser.

### 2.5.2. Posterior predictive p-value

La *posterior predictive p-value* (abrégée désormais en **ppp**) est une des nombreuses GOF p-values existantes et est sûrement à la fois la plus répandue et la plus méconnue. Elle est d'ailleurs souvent nommée à tort « *bayesian p-value* », à tort car en vérité il existe une multitude de p-valeur bayésiennes telles que la *prior predictive p-value* (Box, 1980), la *partial posterior predictive p-value* et la *conditional predictive p-value* (Bayarri & Berger, 2000, 2004), la *sampled posterior p-value* (Gosselin, 2011; Johnson, 2004, 2007) ou encore la *calibrated posterior predictive p-value* (Hjort, Dahl, & Hognadottir, 2006).

Comme nous l'avons dit dans l'introduction, les GOF p-values ont comme ingrédients une fonction de divergence et une loi de réplication de données avec comme objectif de répondre à la question suivante : « Les données observées sont-elles cohérentes avec le modèle ajusté ? ». Pour cela on réplique des données $y_{rep}$ à partir d'une loi de réplication $m(y_{rep}, \theta)$ et on se pose la question si oui

ou non les données répliquées sont compatibles avec le modèle. L'idée des données répliquées est que ce sont des données que l'on aurait pu observer ou que l'on observerait dans le futur si on reconduisait la même expérience qui a produit $y_{obs}$ avec le même modèle et les mêmes valeurs de paramètres. On pose dans toute la suite la notation $\mathbb{P}^{m(V)}(E)$ comme la probabilité de l'évènement $E$ quand la variable aléatoire $V$ a comme distribution de probabilité la fonction $m()$.

Le fonction de divergence notée $T(y, \theta)$ standardise les données et les paramètres en les ramenant à un scalaire : c'est une sorte de résumé. Cette fonction doit être choisie judicieusement de sorte à mesurer des aspects pertinents du modèle étudié. Par exemple si l'on veut détecter un décalage sur le coefficient d'asymétrie (*skewness*), dans ce cas la fonction de divergence la plus adaptée serait la mesure du skewness. Il y a ensuite la loi de réplication des données qui dans le cas de la *posterior predictive p-value* est définie comme :

$$\begin{aligned} m(y_{rep}, \theta | y_{obs}) &= \mathbb{P}(y_{rep}, \theta | y_{obs}) \\ &= \mathbb{P}(y_{rep} | \theta, y_{obs}) \pi(\theta | y_{obs}) \\ &= \mathbb{P}(y_{rep} | \theta) \pi(\theta | y_{obs}) \end{aligned} \quad (1)$$

Notons que la notation ci-dessus est légèrement abusive mais ceci étant pour éviter d'introduire de nouvelles lois de probabilités et de notations. La densité de l'équation (1) est appelée *posterior predictive density* d'où le nom de cette GOF p-value. Le passage de la ligne 2 à la ligne 3 de l'équation (1) est justifié par le fait que, conditionnellement à $\theta$, $y_{obs}$ et $y_{rep}$ sont indépendants.

Donc pour obtenir la ppp il suffit de (i) définir une fonction de divergence $T(y, \theta)$ et (ii) de tirer $N$ fois $\theta_i \in \pi(\theta | y_{obs}), i = \{1, \dots, N\}$ et de générer suivant la loi $m(., \theta_i)$ des données répliquées $y_{rep}^i$ puis de calculer la probabilité que les données répliquées soient plus extrêmes que les données observées du point de vue de la fonction de divergence :

$$ppp = \mathbb{P}^{m(y_{rep}, \theta | y_{obs})}[T(y_{rep}, \theta) > T(y_{obs}, \theta)] \quad (2)$$

La probabilité est ainsi prise sur la distribution *a posteriori* de $\theta$ et sur la *posterior predictive density* de $y_{rep}$, i.e, la distribution jointe de $\theta, y_{rep} | y_{obs}$ :

$$ppp = \int_{\Theta} \int_{\mathbb{R}^n} \chi_{[T(y_{rep}, \theta) > T(y_{obs}, \theta)]} \, p(y_{rep} | \theta) \pi(\theta | y_{obs}) dy_{rep} d\theta \quad (3)$$

où $\chi_A$ représente la fonction indicatrice sur l'ensemble $A$.

La *posterior predictive p-value* est assez directe à programmer mais a un inconvénient majeur. En effet la ppp tend à être conservatrice dans le sens où la distribution de probabilité de cette p-valeur calculée sous l'hypothèse que le modèle de génération des données et le modèle d'estimation sont les mêmes est souvent en forme de dôme (voire piquée) autour de 0.5 plutôt qu'uniforme comme l'est une p-valeur fréquentiste dans les tests d'hypothèses classiques (Robins et al., 2000). En particulier la distribution de la p-value dépend de la fonction de divergence et (Johnson, 2007) a montré que sous certaines conditions la p-value a une distribution uniforme. Ce problème de distribution de probabilité survient car les données sont utilisées deux fois : la première pour calculer la distribution *a posteriori* des paramètres et une deuxième fois pour calculer la probabilité (Bayarri & Berger, 2000). Evidemment, le degré de conservatisme peut varier suivant les données, le modèle et la fonction de divergence, ce qui rend d'autant plus compliquée la comparaison de ppp entre modèles.

Dans un exemple (Zhang, 2014) a montré que la ppp ne rejetait presque jamais le modèle ajusté, même lorsque le modèle utilisé pour ajuster les données diffère complètement de celui utilisé pour générer les données. En outre, (Hjort et al., 2006) a noté que la distribution de la ppp peut prendre des formes très variées. En revanche, une *posterior predictive p-value* proche de 0 ou proche de 1 (par exemple 0.05 ou 0.95) indiquent que le modèle ajusté n'est pas bon.

On l'aura compris, la *posterior predictive p-value* présente un très sérieux inconvénient qui est sa distribution de probabilité incertaine et donc rendant son interprétation quasi impossible. On dispose pourtant une solution qui est d'utiliser une fonction de divergence vérifiant des propriétés de *pivotal quantities* (Johnson, 2007) mais ces propriétés sont trop restrictives sur $T(y,\theta)$ et c'est pourquoi nous introduisons la *sampled posterior predictive p-value*.

### 2.5.3. Sampled posterior predictive p-value

La *posterior predictive p-value* se basait sur un tirage multiple de valeurs de $\theta$ dans la distribution *a posteriori*. Le fonctionnement de la *sample posterior predictive p-value* (abrégée en **sppp**) est le même que la *posterior predictive p-value* à l'exception près qu'on ne tire qu'une seule valeur de $\theta$, notée $\theta_{sampled}$ , d'où le terme *sampled* ! Ainsi pour la *sampled posterior predictive p-value* la loi de réplication des données est :

$$m(y_{rep},\theta|y_{obs}) = p(y_{rep}|\theta)\delta_{\theta_0}(\theta)$$

où $\delta_{\theta_0}(.)$ représente la masse de Dirac en un $\theta_0$ qui est un unique tirage aléatoire dans la distribution *a posteriori* $\pi(\theta|y_{obs})$.

Le fait d'échantillonner $\theta$ semble être un choix préférable en termes d'erreur de type I et de puissance de détection de mauvaise spécification de modèle (Gosselin, 2011; Zhang, 2014). En fait la *sampled posterior predictive p-value* est garantie d'avoir une distribution uniforme sous le modèle nul, i.e, si le modèle ajusté est le même que le modèle de génération des données et en outre la *sampled posterior predictive p-value* suit asymptotiquement une loi uniforme quelle que soit la fonction de divergence (Gosselin, 2011). Ce qui est un énorme atout comparé à la *posterior predictive p-value* qui elle n'avait pas de distribution de référence même sous le modèle nul.

Toutefois, le fait d'utiliser des données simulées et non des données réelles rend réticents de nombreuses personnes familières avec les statistiques, bayésiennes ou non (Piccinato, 2000). En effet la double utilisation des données fait passer les GOF p-values pour une cause perdue où l'on souhaite le beurre – estimer les paramètres avec toutes les données – et l'argent du beurre – avoir des données pour confronter le modèle.

De façon surprenante la *sampled posterior predictive p-value* fournit une p-value uniforme malgré la double utilisation des données. Par conséquent elle semble avoir exactement le même défaut d'utilisation double des données que la *posterior predictive p-value.* Défaut qui était supposé justifier son absence de distribution. La preuve faite dans (Gosselin, 2011) révèle une autre explication : l'échantillonnage de $\theta$. Ce choix peut sembler étrange car cela rend la *sampled posterior predictive p-value* aléatoire, dans le sens où à même données et même modèle on peut obtenir deux valeurs différentes de *sampled posterior predictive p-value.* Mais comme (Gosselin, 2011) le signale : « comme nous travaillons sur des données échantillonnées pour ajuster des modèles statistiques, nous devrions être d'accord pour travailler sur des paramètres échantillonnés pour critiquer le(s) modèle(s). En effet ce double échantillonnage nous permet de symétriser le rôle des données et des

paramètres. »[1]. De la même façon que des données échantillonnées peuvent avoir une faible probabilité conditionnelle aux conditions réelles du vrai modèle n'empêchent par leur utilisation dans l'ajustement de modèles, le fait que les paramètres échantillonnés peuvent avoir une faible probabilité conditionnelle aux données n'empêche pas leur utilisation dans la vérification de modèles. Ainsi le problème n'est pas tant le fait que les GOF p-values utilisent deux fois les données que **comment** elles sont utilisées – voir (Evans, 2007).

Maintenant que nous avons posé le cadre statistique dans lequel se place ce stage ainsi que les outils utilisés, nous allons pouvoir passer au protocole utilisé pour le stage.

[1] « as we are working on sampled data to fit statistical models, we should also agree to work on sampled parameters to criticize the model. Indeed, this double sampling allowed us to make the roles of data and parameters symmetrical. »(Gosselin, 2011)

## 2.6. Approche adoptée

Tout d'abord, commençons par rappeler que nous allons adopter une démarche de simulation pour avoir un regard plus objectif sur les résultats. En effet, en connaissant les vraies valeurs des paramètres nous allons pouvoir comparer celles-ci avec les valeurs estimées en toute objectivité. Ainsi grâce au code de (Coron et al., 2018) que nous avons modifié nous allons :

i. Générer des données
ii. Ajuster le modèle statistique de (Coron et al., 2018) via notamment l'utilisation d'un MCMC à une seule chaîne de Markov.
iii. Appliquer la sampled posterior predictive p-value

Les deux premières étapes décrites ci-dessus sont itérées en boucle « for » et à chaque passage dans la boucle on enregistre un fichier RData qui servira dans la partie GOF p-values. De plus, en s'appuyant sur les hypothèses faites par le modèle statistique statistique de (Coron et al., 2018), nous avons élaboré des scénarii de simulation – donc de modèle probabiliste – afin de voir ce que la sampled posterior predictive p-value permet de détecter ou non en terme d'inadéquation entre le modèle statistique et le modèle probabiliste. Nous allons commencer par détailler le fonctionnement de la boucle while du MCMC et puis nous exposerons les détails des scénarii. Aussi, contrairement à (Coron et al., 2018), nous ne souhaiterions pas utiliser trois jeux de données mais un seul, constitué de données opportunistes et de données protocolées, et qui nous sert à l'estimation statistique. Ensuite le but est de vérifier la cohérence entre données opportunistes et données protocolées via les GOF p-values.

### 2.6.1. Boucle while et GOF

La boucle while commence après (i) la génération des données et (ii) après un premier calcul de 10 000 itérations MCMC avec un thin=50 et n.adapt=1000. Le but étant d'avoir un MCMC convergent, i.e, un MCMC post burn-in et avec un thin adéquat. Or nous avons constaté que cette partie d'étude de convergence était manquante dans le code de (Coron et al, 2018.). De plus, lors des premiers essais de code, la partie MCMC posait de gros problèmes. En effet, le burn-in estimé (environ 10000) et le thin estimé (environ 500) les premières fois nous semblaient sur-estimés par rapport aux graphiques de trajectoires de chaînes que l'on a tracé. C'est donc là tout l'intérêt de cette boucle while : assurer de bonnes estimations du burn-in et du thin de sorte à avoir un MCMC convergent tout en n'ayant pas à faire 1 million d'itérations pour en avoir 100 convergentes.

Tout d'abord dans cette boucle on fait une première estimation du burn-in en utilisant le score de Geweke récursivement et en enlevant des itérations au fur et à mesure. Ceci terminé, on calcule une première fois le thin en prenant en compte le burn-in précédemment calculé. On a donc une première estimation du burn-in et du thin, ce qui nous permet de déterminer le nombre d'itérations totales à faire faire par le MCMC pour en avoir le nombre souhaité convergentes.

Ensuite on va récursivement :

i. Continuer le MCMC en faisant une partie du nombre total d'itérations estimé à faire.
ii. Calcul du burn-in via le score de Geweke utilisé récursivement en enlevant les premières itérations au fur et à mesure.
iii. Calcul du thin en prenant en compte le burn-in estimé à l'étape précédente.
iv. Mise à jour du nombre total d'itérations à faire en prenant en compte le burn-in et le thin que l'on vient d'obtenir par les deux étapes précédentes.

A l'issue de cette boucle on enregistre un fichier RData. Ensuite vient l'analyse des fichiers RData via les GOF p-values. Nous nous sommes alors demandés combien de fichier RData il nous fallait, compte tenu de la contrainte du temps de calcul et du temps de stage restant. En effet il faut environ 1h15 pour obtenir un RData et 6h pour faire les GOF p-values sur 10 RData. Sachant que l'on a 5 scénarii, que l'on abordera juste après, cela donne environ 295 jours par scénario si l'on souhaite obtenir et analyser 100 RData… Nous avons donc dû utiliser le cluster disponible à Nogent-Sur-Vernisson et faire tourner en parallèle plusieurs session de RStudio pour mener à bien tous les calculs en temps voulu. Nous avons cherché un bon équilibre sur le nombre de RData à considérer entre temps de calcul trop long et bonne estimation des GOF p-values. Le fruit de cette considération est disponible en annexe de ce document et nous sommes arrivés au choix de 300 RData.

L'étape suivante est le diagnostic par GOF p-values. Cette étape est itérée sur les RData. Le fonctionnement est le suivant :

i. Charger un RData.
ii. Tirer au sort des observations de sorte à ne pas toutes les prendre. Sinon la sampled posterior predictive p-value serait « trop » puissante et l'on détecterait des choses trop petites comme par exemple une moyenne à 0.001 au lieu de 0. Ainsi on tire au sort de façon à ce que par site et par espèce l'on ait 2 données protocolées et 6 opportunistes. Ce qui donne $2 \times I \times J + 6 \times I \times J$ données considérées au lieu de $10 \times I \times J + 30 \times I \times J$.
iii. On a grâce au RData les itérations MCMC post burn-in et thinnées. On tire alors au sort une itération car on utilise la sampled posterior predictive p-value, qui par construction fonctionne avec une seule valeur échantillonnée des paramètres. On obtient alors une valeur des paramètres.
iv. Simuler 1000 jeux de données : ce sont les réplications de données.
v. Normaliser les données observées et celles répliquées de sorte qu'elles soient comparables entre elles. En effet, les moyenne, variance, etc des données opportunistes et des données protocolées ne sont pas forcément sur les mêmes échelles de grandeur. On a ainsi des données non normalisées et des données normalisées.
vi. Appliquer les fonctions de discrépances aux données répliquées et aux données observées.
vii. Calculer la p-valeur en la tirant d'une loi Beta de paramètres $\alpha + 1$ et $\beta + 1$ où

$$\alpha = \sum_j 1_{T\left(y_{rep_j}, \theta_{sampled}\right) > T\left(y_{obs_j}, \theta_{sampled}\right)} + \epsilon \sum_j 1_{T\left(y_{rep_j}, \theta_{sampled}\right) = T\left(y_{obs_j}, \theta_{sampled}\right)}$$

$$\beta = \sum_j 1_{T\left(y_{rep_j}, \theta_{sampled}\right) < T\left(y_{obs_j}, \theta_{sampled}\right)} + (1 - \epsilon) \sum_j 1_{T\left(y_{rep_j}, \theta_{sampled}\right) = T\left(y_{obs_j}, \theta_{sampled}\right)}$$

et $\epsilon$ est tiré d'une loi Uniforme sur l'intervalle [0,1] (Gosselin, 2011). En effet on peut montrer que la p-valeur sous-jacente est issue de cette distribution. Ne pas tirer la p-valeur dans cette distribution pourrait amener à des écarts significatifs de la distribution uniforme. De plus nous transformons la p-valeur $p$ via la fonction $p \longmapsto 2\min(p, 1-p)$ pour concentrer les p-valeurs surprenantes autour de 0 au lieu d'être autour de 0 et/ou 1.

On obtient alors une p-valeur par RData. On rappelle que la sampled posterior predictive p-value est issue d'une loi uniforme sur [0,1] si les données sont générées suivant le même modèle que celui ajusté, ou dit autrement si le modèle probabiliste est le même que le modèle statistique. Ainsi, pour pouvoir interpréter nos p-valeurs, il nous faut les comparer à une loi uniforme sur [0,1]. Nous allons donc utiliser un test de Kolmogorov-Smirnov et tracer les histogrammes des p-valeurs. De plus nous sortirons les proportions de p-valeur inférieures à 0.001, 0.01 et 0.05. Toutefois, le calcul de proportions de p-valeur inférieures à 0.001 n'est pas totalement pertinent car nous n'avions que 1000 réplications de données pour la partie GOF. Ce calcul de proportion serait plus pertinent avec au moins 10 000 réplications.

En détails, notre diagnostic pour juger si une fonction de discrépance « accepte » ou « rejette » est le suivant : à chaque RData analysé nous avons une valeur de sampled posterior predictive p-value. Sachant que cette valeur, si les données sont cohérentes avec le modèle statistique (i.e. le modèle probabiliste est le même que le modèle statistique), est issue d'une loi Uniforme sur [0,1]. Donc si nous analysons 300 RData alors nous avons potentiellement 300 réalisations d'une loi Uniforme sur [0,1]. Ainsi, si la p-valeur du test de Kolmogorov-Smirnov n'est pas trop extrême (<5%) alors cela indique que les 300 valeurs sont issues d'une loi Uniforme sur [0,1]. En remontant le fil, cela indique que les fonctions de discrépance ne détectent pas de problème et donc que le modèle statistique est cohérent et bien ajusté aux données.

### 2.6.2. Scénarii, codes R associés et fonctions de discrépances

Maintenant que nous avons vu la méthode utilisée pour simuler les données et les analyser avec les GOF p-values, nous allons détailler les scénarii de simulation. En effet comme nous l'avons dit précédemment nous allons changer certaines hypothèses faites par (Coron et al., 2018) afin de voir ce qui est détecté ou non par notre méthode. Nous allons également exposer les fonctions de discrépances utilisées. Celles-ci sont divisées en deux groupes. Le premier appelé « omnibus » regroupe les fonctions de discrépances qui ne ciblent pas un comportement particulier comme par exemple la moyenne ou la variance. Le deuxième groupe appelé « ciblées » est constitué de fonctions qui ciblent un caractère bien précis en lien avec une violation d'hypothèse. Les fonctions de discrépances utilisées sont regroupées dans un tableau à la fin de cette sous-section. En outre, pour les scénarii à l'exception du scénario 1, nous avons choisi (i) des variations réalistes avec graphiques à l'appui et (ii) des hypothèses qu'on retrouve dans la littérature ou qui semblent intéressantes d'un point de vue écologique ou pratique

#### Scénario 1 : Données générées suivant le modèle statistique

Le premier scénario est en quelque sorte un scénario témoin. Il s'agit de générer et d'analyser les données avec exactement le même modèle. Dit autrement on va faire correspondre le modèle probabiliste avec le modèle statistique. Pour faire une analogie avec un modèle de régression linéaire, c'est comme si on générait des données iid, gaussiennes avec un bruit homoscédastique et que l'on y appliquait une régression linéaire. Ce faisant, les données vérifieraient exactement les bons postulats pour pouvoir appliquer une régression linéaire nous garantissant alors un modèle excellent. Ainsi, en vérifiant les résidus on aurait tous les voyants au vert (résidus uniformes, homoscédastiques etc). Pour revenir à notre modèle principal, on s'attend donc dans ce premier scénario à ce que les fonctions de discrépances ne détectent rien.

Quant au code R de simulation des données, qui est celui de (Coron et al, 2018.), le processus est le suivant : on pose les constantes comme le nombre d'espèces, le nombre de sites etc. Puis on tire au sort les paramètres du modèle tels que la probabilité de reporting, la sélection d'habitat des espèces etc. Enfin on génère un dataset en tirant dans une loi de Poisson ayant comme paramètre celui de (Coron et al, 2018.) : $N_{ij}E_{ck}P_{ik}\sum_h \frac{q_{hk}}{\sum_{h'} q_{h'k}V_{h'c}} \times \frac{S_{ih}}{\sum_{h'} S_{ih'}V_{h'j}} V_{hc}$.

Pour ce scénario, nous prendrons toutes les fonctions de discrépances dont nous nous servirons par la suite. En effet le but est de contrôler que ces fonctions ne détectent rien dans ce scénario !

#### Scénario 2 : Dépendance spatiale pour la sélection d'habitat des espèces

Dans ce scénario nous allons modifier la probabilité de sélection de l'habitat $h$ par l'espèce $i$. En effet cette probabilité était uniforme et ne dépendait pas du site $j$ dans le scénario 1. C'est pourquoi nous modifions sa définition dans le code R ainsi que son influence dans X.

Nous faisons alors apparaître une dépendance à une échelle locale dans la probabilité de sélection d'habitat par les espèces. Ce code permet par exemple de passer d'un ratio de 1 pour 3 à un ratio de 3 pour 1 sur certaines cellules, ce que ne permettait pas le code du scénario 1.

### Scénario 3 : préférence d'habitat non uniforme pour les données opportunistes

Une hypothèse du scénario 1 était que les préférences d'habitats pour les données opportunistes étaient constantes entre sites. Or on pourrait faire varier les préférences suivant les sites $j$, de la même façon que l'on avait fait varier les préférences d'habitats pour les espèces dans le scénario 2. Ainsi on construit $q$ de sorte que ses dimensions soient $H \times K \times J$ et non plus juste $H \times K$ comme dans le scénario 1.

### Scénario 4 : dépendance entre probabilité de reporting et préférence de l'espèce pour un habitat

Ce scénario consiste à introduire une dépendance entre la probabilité de reporting d'une espèce notée $P_{ik}$ et sa préférence pour un habitat notée $S_{ih}$. On ajoute alors une dimension à la probabilité de reporting correspondant à l'habitat. Ainsi cette probabilité a désormais pour dimensions $I \times H \times K$ puis on multiplie, à $k$ fixé, la probabilité de reporting par la préférence d'habitat des espèces. Et pour finir, à $k$ fixé, on somme sur $h$. On retrouve alors une probabilité de reporting $P_{ik}$ à dimensions $I \times K$.

### Scénario 5 : abondance sous-jacente $N_{ij}$ variable inter-cellules

Comme nous l'avons vu dans la section évoquant les hypothèses du code et du modèle statistique, une hypothèse forte est que l'abondance sous-jacente $N_{ij}$ ne dépend pas de la cellule $c$. Dans ce scénario, nous allons ajouter cette dépendance à la cellule afin d'intégrer dans l'abondance sous-jacente une variabilité **intra**-site en plus de la variabilité **inter**-sites présente dans le modèle originel.

Dans ce scénario on crée Nr0 avec la même définition que le Nr du scénario 1 puis on modifie ce Nr0 lui appliquant une transformation. La Figure 2 montre l'effet de cette transformation.

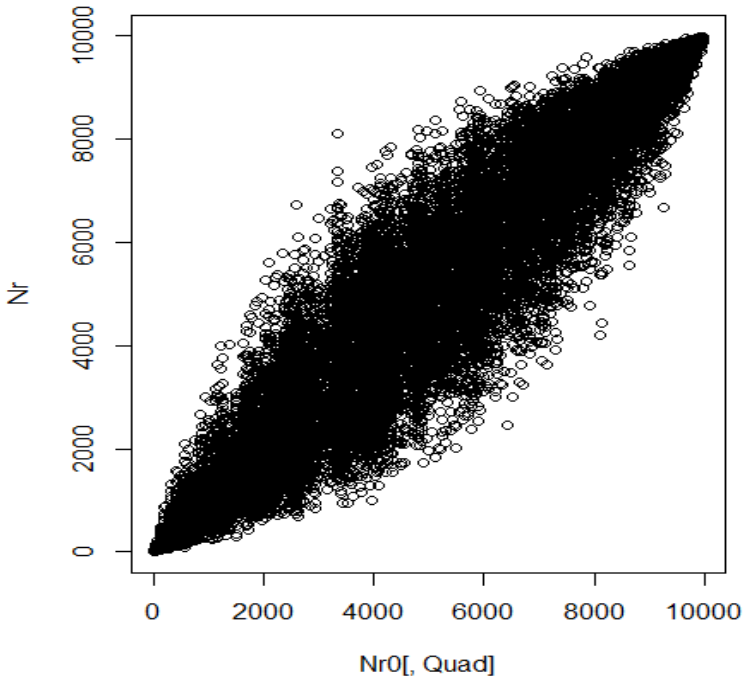


*Figure 2: Nr0 en fonction de Nr*

Le Tableau 1 récapitule les fonctions de discrépances utilisées suivant chaque scénario. En lignes sont disposées les fonctions de discrépances et en colonnes ce sont les scénarii. De plus pour les fonctions de discrépances utilisant *l'Akaike Information Criterion* (AIC) nous utilisons la syntaxe de R. A savoir AIC(eff~site*esp) signifie par exemple que nous calculons l'AIC du modèle linéaire qui a pour variable à expliquer l'abondance (eff pour effectif) et qui a pour variable explicative l'interaction entre la cellule et l'espèce pris comme facteurs. Pour prendre un autre exemple, AIC(eff~1) signifie que l'on calcule l'AIC du modèle linéaire qui explique la variable « eff » par un effet constant, correspondant au « 1 » de « ~1 ». Le calcul de l'AIC sous R est : -2*log-likelihood+2*nbparametres.

Remarquons que les cinq premières fonctions de discrépances du Tableau 1 sont les discrépances « omnibus » et les cinq dernières sont les discrépances « ciblées ». De plus, si une case est colorée en rouge alors la fonction de discrépance qui est sur la ligne est utilisée pour le scénario de la colonne correspondante. Par exemple, d'après le Tableau 1, pour le scénario 2 nous utiliserons les fonctions de discrépances suivantes : moyenne, variance, skewness, kurtosis et AIC(eff~site*esp)-AIC(eff~site+esp). On utilise ainsi dix fonctions de discrépance et dans la partie des résultats nous nous référerons aux fonctions de discrépance suivant leur numéro, de un à dix par ordre descendant du Tableau 1.

*Tableau 1 : fonctions de discrépances utilisées suivant chaque scénario*

| | Scénario 1 | Scénario 2 | Scénario 3 | Scénario 4 | Scénario 5 |
|---|---|---|---|---|---|
| Moyenne | ■ | ■ | ■ | ■ | ■ |
| Variance | ■ | ■ | ■ | ■ | ■ |
| Kurtosis | ■ | ■ | ■ | ■ | ■ |
| Skewness | ■ | ■ | ■ | ■ | ■ |
| Variance/Moyenne | ■ | ■ | ■ | ■ | ■ |
| AIC(eff~site*esp) - AIC(eff~site+esp) | ■ | ■ | ■ | | |
| AIC(eff~dataset+esp)-AIC(eff~1) | ■ | | | ■ | |
| AIC(eff~dataset*esp)-AIC(eff~1) | ■ | | | ■ | |
| AIC(eff~site)-AIC(eff~1) | ■ | | | | ■ |
| AIC(eff~site*esp)-AIC(eff~1) | ■ | | | | ■ |

Pour finir nous allons aborder les modifications que nous avons apportées au code utilisé par (Coron et al., 2018). Nous ne pouvions pas évoquer ces modifications plus haut car celles-ci se justifient avec les fonctions de discrépances (cf Tableau 1).

En premier lieu, nous avons ajouté en facteur multiplicatif dans le modèle statistique de JAGS le terme suivant :

$$\frac{1}{\sum_{h'} q_{h'k} V_{h'c} \times \sum_{h'} S_{ih'} V_{h'j}}$$

Et ceci pour faire correspondre le modèle probabiliste (modèle qui génère les données) avec le modèle statistique (modèle qui estime des paramètres à partir des données). De plus, dans un contexte appliqué, nous avons accès à ces quantités écologiques, il est ainsi préférable de les ajouter au modèle pour augmenter la qualité d'estimation. En effet, dans le code originel utilisé on n'utilisait que les termes suivants :

$$N_{ij} E_{ck} P_{ik} \sum_{h} q_{hk} \times S_{ih} \times V_{hc}$$

Passés en paramètres de la loi de Poisson. Ainsi, nous utilisons comme paramètre de la loi de Poisson sous JAGS le terme :

$$N_{ij} E_{ck} P_{ik} \sum_{h} \frac{q_{hk}}{\sum_{h'} q_{h'k} V_{h'c}} \times \frac{S_{ih}}{\sum_{h'} S_{ih'} V_{h'j}} V_{hc}$$

Ce qui correspond exactement au modèle exposé dans (Coron et al., 2018).

Le deuxième changement majeur se situe dans les lois *a priori*. Nous avons choisi des lois normales avec une forte variance valant 25 sur une échelle logarithmique et non des lois uniformes sur l'intervalle [0,100] comme l'a fait (Coron et al., 2018).

Le choix de prendre une loi normale à forte variance est une façon répandue d'utiliser une loi *a priori* à faible contenu informatif. C'est-à-dire que cette loi *a priori* n'influence pas trop les valeurs des paramètres en comparaison avec une loi qui serait très piquée en une valeur comme une loi normale de variance 1/2. Dit autrement, ce choix de loi *a priori* permet aux paramètres de prendre des valeurs extrêmes car avec des lois uniformes il y avait des problèmes de bornes. Une variance élevée permet d'éviter que les GOF ne détecte des problèmes à cause de la loi *a priori* et non à cause des données. Le passage à l'exponentielle se justifie écologiquement. L'espace d'état du paramètre $N_{ij}, 1 \leq i \leq I\ 1 \leq j \leq J$ est logarithmique pour faire en sorte qu'une multiplication par une constante de l'abondance ait la même probabilité qu'une division par cette même constante.

Tous ces changements sont également justifiés par la sampled posterior predictive p-value. Nous avons pris exactement le même code que (Coron et al., 2018) pour la partie statistique et avons sorti 10 RData en vue de leur appliquer la sampled posterior predictive p-value en utilisant le même procédé que décrit plus haut. Le problème majeur était que toutes les fonctions de discrépances omnibus étaient rejetées très franchement avec une p-valeur inférieure à $2.2^{e}$-16, quand elles étaient appliquées sur les données non normalisées. En outre la normalisation des données de sorte à les comparer à une loi normale centrée-réduite ne renvoyait que des Nan ce qui indique un problème d'extrémité dans les données.

# 3. Résultats

Dans cette partie nous allons exposer les résultats obtenus à l'issue du protocole détaillé précédemment. Nous ferons une section par scénario. On rappelle que pour obtenir ces résultats nous avons analysé 300 RData par scénario et pour cela nous avons utilisé un ordinateur avec un processeur Intel® Core™ i7-6820HQ 2.70GHz et 16Go de RAM. De plus par contrainte de temps nous n'avons pas directement analysé les 300 RData mais avons coupé chaque scénario en 6 processus de 50 RData. Le temps de calcul d'un processus varie d'un scénario à l'autre et sera indiqué dans la section du scénario correspondant. De plus les calculs ont été fait en singlecore et non en HPC.

## 3.1. Scénario 1

Pour ce premier scénario, chaque processus de 50 RData a mis 78 heures soit 3.25 jours. On voit grâce au Tableau 2 que l'on accepte toutes les fonctions de discrépances sur données non normalisées. Cela se confirme avec les histogrammes de la Figure 3. En revanche il y a certaines fonctions de discrépances qui rejettent sur les données normalisées, en particulier la 8 qui est très nette. Cela montre alors l'intérêt de la normalisation des données, et cela s'explique en partie par le fait que la fonction de discrépance 8 utilise un modèle linéaire qui est donc plus performant avec des données issues de loi normale. Et étant donné que l'on utilise que des modèles linéaires pour les fonctions de discrépance 5 à 10, cela explique la significativité plus forte sur les données normalisées.

Ainsi la sampled posterior predictive p-value ici ne détecte globalement pas de problèmes ce qui est normal car il n'y en a pas : nous avons fait correspondre modèle probabiliste et modèle statistique.

*Tableau 2 : résultats du test de Kolmogorov-Smirnov sur scénario 1*

| | Données non normalisées | Données normalisées |
|---|---|---|
| Moyenne | Statistique :0.04<br>p-valeur : 0.70 | Statistique :0.05<br>p-valeur : 0.33 |
| Variance | Statistique :0.03<br>p-valeur : 0.93 | Statistique :0.04<br>p-valeur : 0.65 |
| Kurtosis | Statistique :0.04<br>p-valeur : 0.68 | Statistique :0.04<br>p-valeur : 0.55 |
| Skewness | Statistique :0.07<br>p-valeur : 0.10 | Statistique :0.05<br>p-valeur : 0.37 |
| Variance/Moyenne | Statistique :0.04<br>p-valeur : 0.48 | Statistique :0.05<br>p-valeur : 0.49 |
| AIC(eff~site*esp) -<br>AIC(eff~site+esp) | Statistique :0.05<br>p-valeur : 0.29 | Statistique :0.11<br>p-valeur : 0.0007022 |
| AIC(eff~dataset+esp)-AIC(eff~1) | Statistique :0.05<br>p-valeur : 0.36 | Statistique :0.07<br>p-valeur : 0.12 |
| AIC(eff~dataset*esp)-AIC(eff~1) | Statistique :0.05<br>p-valeur : 0.36 | Statistique :0.18<br>p-valeur : 2.183e-09 |
| AIC(eff~site)-AIC(eff~1) | Statistique :0.05<br>p-valeur : 0.45 | Statistique :0.04<br>p-valeur : 0.57 |
| AIC(eff~site*esp)-AIC(eff~1) | Statistique :0.03<br>p-valeur : 0.90 | Statistique :0.12<br>p-valeur : 0.0001503 |

*Tableau 3 : Proportion de p-valeurs inférieures à 0.001, 0.01 et 0.05 pour le scénario 1*

| | Proportion p-valeur<0.001 | | Proportion p-valeur<0.01 | | Proportion p-valeur<0.05 | |
|---|---|---|---|---|---|---|
| Moyenne | Norm : 0.00333 | Non norm : 0.006666 | Norm : 0.01666 | Non norm : 0.01333 | Norm : 0.07666 | Non norm : 0.05666 |
| Variance | Norm : 0.00333 | Non norm : 0 | Norm : 0.01333 | Non norm : 0.01333 | Norm : 0.05666 | Non norm : 0.06 |
| Kurtosis | Norm : 0 | Non norm : 0 | Norm : 0.006666 | Non norm : 0.01333 | Norm : 0.05 | Non norm : 0.05333 |
| Skewness | Norm : 0 | Non norm : 0 | Norm : 0.01333 | Non norm : 0.01 | Norm : 0.053 | Non norm : 0.06333 |
| Variance/Moyenne | Norm : 0 | Non norm : 0 | Norm : 0.00666 | Non norm : 0.02 | Norm : 0.046666 | Non norm : 0.0666 |
| AIC(eff~site*esp)-AIC(eff~site+esp) | Norm : 0 | Non norm : 0 | Norm : 0.0233 | Non norm : 0.01 | Norm : 0.09333 | Non norm : 0.05666 |
| AIC(eff~dataset+esp)-AIC(eff~1) | Norm : 0 | Non norm : 0 | Norm : 0.00666 | Non norm : 0.02 | Norm : 0.06666 | Non norm : 0.09 |
| AIC(eff~dataset*esp)-AIC(eff~1) | Norm : 0.01 | Non norm : 0 | Norm : 0.07666 | Non norm : 0.1666 | Norm : 0.18 | Non norm : 0.10333 |
| AIC(eff~site)-AIC(eff~1) | Norm : 0 | Non norm : 0 | Norm : 0.02 | Non norm : 0.01 | Norm : 0.05333 | Non norm : 0.04333 |
| AIC(eff~site*esp)-AIC(eff~1) | Norm : 0.01 | Non norm : 0.003333 | Norm : 0.04 | Non norm : 0.02333 | Norm : 0.09 | Non norm : 0.06333 |

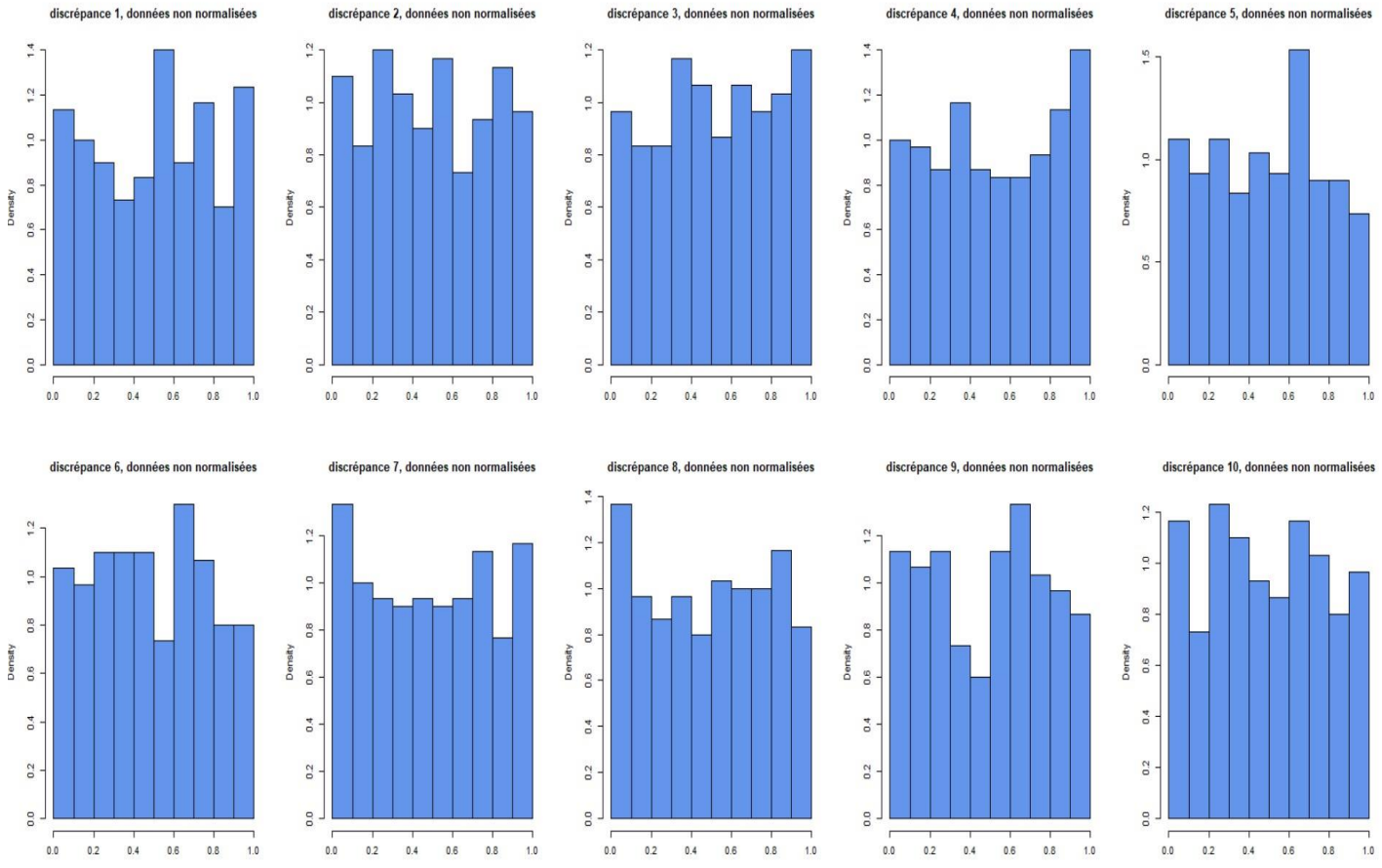


*Figure 3 : histogrammes du scénario 1 sur données non normalisées*

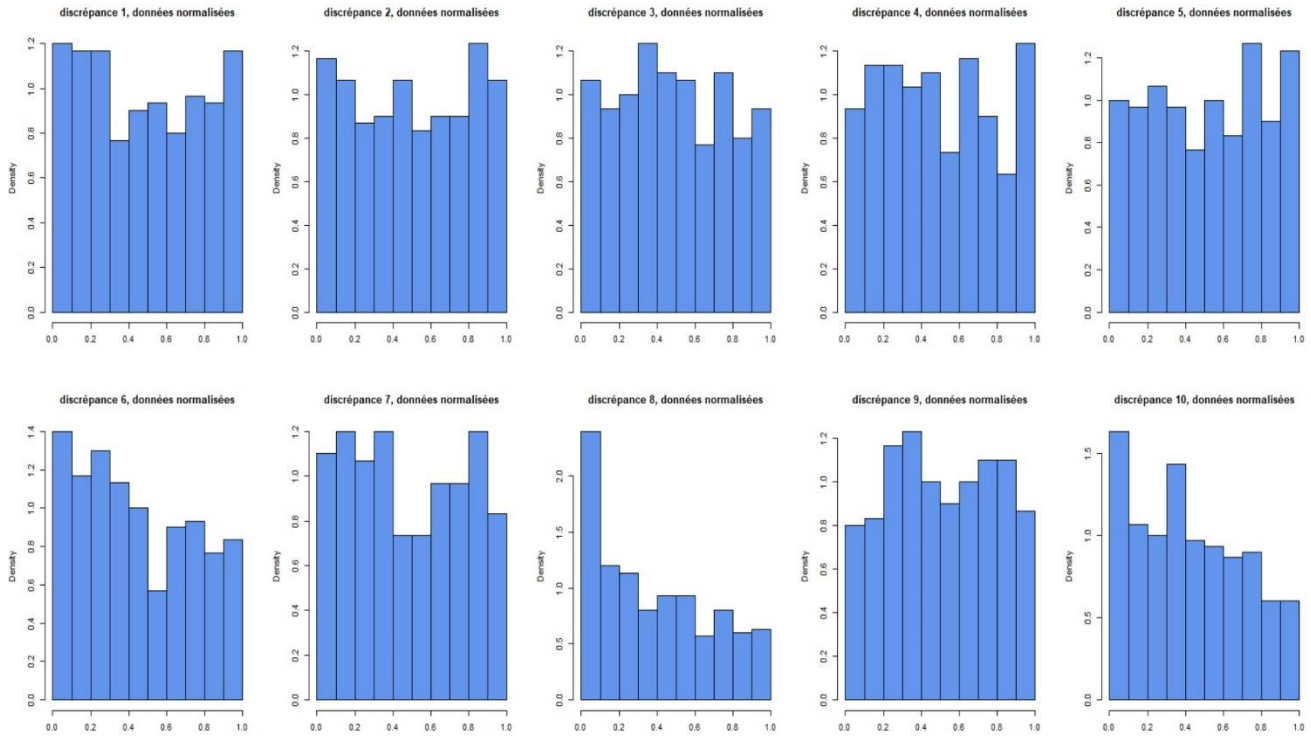


*Figure 4 : histogrammes du scénario 1 sur données normalisées*

Le

Tableau 3 montre que globalement, les proportions ne sont pas choquantes et ne poussent pas à

| | Proportion p-valeur<0.001 | | Proportion p-valeur<0.01 | | Proportion p-valeur<0.05 | |
|---|---|---|---|---|---|---|
| Moyenne | Norm : 0.00333 | Non norm : 0.006666 | Norm : 0.01666 | Non norm : 0.01333 | Norm : 0.07666 | Non norm : 0.05666 |
| Variance | Norm : 0.00333 | Non norm : 0 | Norm : 0.01333 | Non norm : 0.01333 | Norm : 0.05666 | Non norm : 0.06 |
| Kurtosis | Norm : 0 | Non norm : 0 | Norm : 0.006666 | Non norm : 0.01333 | Norm : 0.05 | Non norm : 0.05333 |
| Skewness | Norm : 0 | Non norm : 0 | Norm : 0.01333 | Non norm : 0.01 | Norm : 0.053 | Non norm : 0.06333 |
| Variance/Moyenne | Norm : 0 | Non norm : 0 | Norm : 0.00666 | Non norm : 0.02 | Norm : 0.046666 | Non norm : 0.0666 |
| AIC(eff~site*esp)-AIC(eff~site+esp) | Norm : 0 | Non norm : 0 | Norm : 0.0233 | Non norm : 0.01 | Norm : 0.09333 | Non norm : 0.05666 |
| AIC(eff~dataset+esp)-AIC(eff~1) | Norm : 0 | Non norm : 0 | Norm : 0.00666 | Non norm : 0.02 | Norm : 0.06666 | Non norm : 0.09 |
| AIC(eff~dataset*esp)-AIC(eff~1) | Norm : 0.01 | Non norm : 0 | Norm : 0.07666 | Non norm : 0.1666 | Norm : 0.18 | Non norm : 0.10333 |
| AIC(eff~site)-AIC(eff~1) | Norm : 0 | Non norm : 0 | Norm : 0.02 | Non norm : 0.01 | Norm : 0.05333 | Non norm : 0.04333 |
| AIC(eff~site*esp)-AIC(eff~1) | Norm : 0.01 | Non norm : 0.003333 | Norm : 0.04 | Non norm : 0.02333 | Norm : 0.09 | Non norm : 0.06333 |

rejeter les fonctions de discrépances exceptées les fonctions 7 et 8. En effet, si la proportion de p-valeur inférieure à respectivement 0.001, 0.01 et 0.05 ne sont pas significativement plus grande que 0.001, 0.01 et 0.05 alors c'est que les p-valeurs sont bien issues d'une loi Uniforme sur [0,1]. Sauf

bien entendu pour le cas de la fonction pathologique variance/moyenne comme nous allons le voir dans les autres scénarios.

Ainsi nous pouvons avoir des éléments de comparaison pour les autres scénarii et on s'attend à ce que toutes les fonctions de discrépance rejettent très nettement comme le fait la discrépance 8 sur données normalisée pour ce premier scénario.

## 3.2. Scénario 2

Pour ce deuxième scénario, chaque processus de 50 RData a mis 12 heures. Le test de Kolmogorov-Smirnov est déjà très clair : on rejette franchement toutes les fonctions de discrépances exceptée la première sur données non normalisées. Cela se retrouve sur les figures 5 et 6. En effet les histogrammes sont piqués en 0, indiquant un problème (i.e. une inadéquation du modèle statistique par rapport aux données) fort sur presque tous les jeux de données. On peut remarquer aussi le comportement qu'a la cinquième fonction de discrépance sur les données normalisées qui ressemble à une loi en courbe de cloche. Bien que ce comportement ne soit pas l'un de ceux attendus (soit piqué en 0 soit histogramme plat) cela montre bien un problème. On suppose que cela vient du fait que la moyenne et la variance sont deux quantités piquées en 0 et donc leur rapport a un comportement pathologique.

Ainsi la sampled posterior predictive p-value a détecté des problèmes dans le modèle statistique nous faisant alors comprendre qu'il est à améliorer. De plus la significativité est très forte et on la voit à travers la p-valeur très forte du test de Kolmogorov-Smirnov d'une part mais aussi graphiquement avec les histogrammes très piqués en 0 d'autre part. Cela apparaît aussi à travers le Tableau 5 car les proportions sont significativement plus grande que celles attendues. Par exemple pour la variance sur données normalisées, on observe une proportion de p-valeur inférieure à 0.05 de 0.829 !

*Tableau 4 : résultats du test de Kolmogorov-Smirnov sur scénario 2*

| | Données non normalisées | Données normalisées |
|---|---|---|
| Moyenne | Statistique :0.04<br>p-valeur : 0.75 | Statistique :0.68<br>p-valeur : < 2.2e-16 |
| Variance | Statistique :0.78<br>p-valeur : < 2.2e-16 | Statistique :0.99<br>p-valeur : < 2.2e-16 |
| Kurtosis | Statistique :0.70<br>p-valeur : < 2.2e-16 | Statistique :0.98<br>p-valeur : < 2.2e-16 |
| Skewness | Statistique :0.69<br>p-valeur : < 2.2e-16 | Statistique :0.82<br>p-valeur : < 2.2e-16 |
| Variance/Moyenne | Statistique :0.81<br>p-valeur : < 2.2e-16 | Statistique :0.22<br>p-valeur : 5.982e-09 |
| AIC(eff~site*esp) - AIC(eff~site+esp) | Statistique :0.88<br>p-valeur : < 2.2e-16 | Statistique :0.93<br>p-valeur : < 2.2e-16 |

*Tableau 5 : Proportion de p-valeurs inférieures à 0.001, 0.01 et 0.05 pour le scénario 2*

| | Proportion p-valeur<0.001 | | Proportion p-valeur<0.01 | | Proportion p-valeur<0.05 | |
|---|---|---|---|---|---|---|
| Moyenne | Norm : 0.130 | Non norm : 0 | Norm : 0.485 | Non norm : 0.00975 | Norm : 0.705 | Non norm : 0.0439 |
| Variance | Norm : 0.365 | Non norm : 0.30 | Norm : 1 | Non norm : 0.76585 | Norm : 1 | Non norm : 0.829 |
| Kurtosis | Norm : 0.350 | Non norm : 0.2487 | Norm : 0.99 | Non norm : 0.6487 | Norm : 1 | Non norm : 0.7463 |
| Skewness | Norm : 0.305 | Non norm : 0.234 | Norm : 0.815 | Non norm : 0.6536 | Norm : 0.865 | Non norm : 0.7365 |
| Variance/Moyenne | Norm : 0 | Non norm : 0.321 | Norm : 0 | Non norm : 0.7853 | Norm : 0.035 | Non norm : 0.8536 |
| AIC(eff~site*esp)-AIC(eff~site+esp) | Norm : 0.430 | Non norm : 0.3170 | Norm : 0.940 | Non norm : 0.8975 | Norm : 0.96 | Non norm : 0.917 |

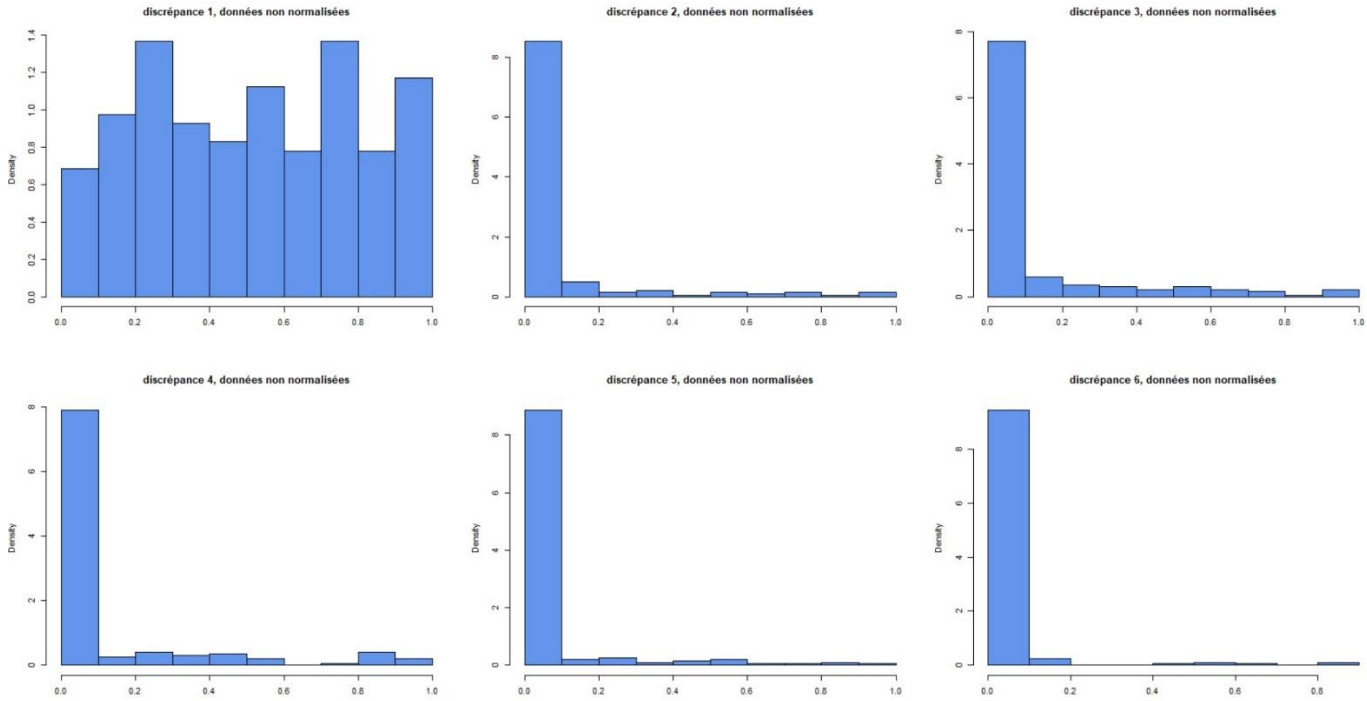


*Figure 5 : histogrammes du scénario 2 sur données non normalisées*

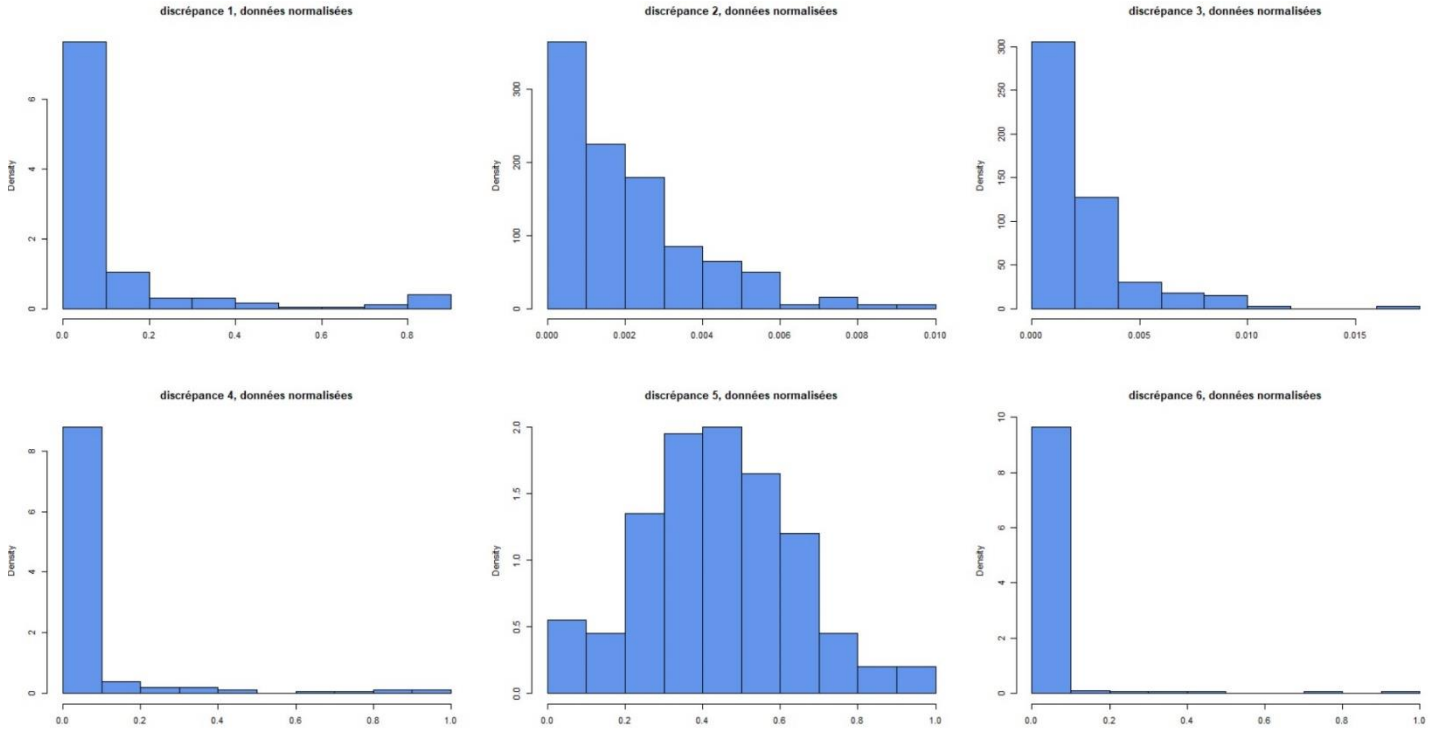


*Figure 6 : histogrammes du scénario 2 sur données normalisées*

## 3.3. Scénario 3

Pour ce troisième scénario, chaque processus de 50 RData a mis 48 heures. Encore une fois le test de Kolmogorov-Smirnov est clair : on rejette franchement toutes les fonctions de discrépances exceptée la première. Cela se retrouve sur les figures 7 et 8. En effet les histogrammes sont piqués en 0, indiquant un problème. Toutefois, il peut sembler sur la Figure 7 que la fonction de discrépance 5 soit non significative car l'histogramme semble plutôt plat. Ceci est infirmé par la valeur de la statistique de test de Kolmogorov-Smirnov qui vaut 0.31 ce qui est du même ordre de grandeur que celle de la fonction de discrépance 3 par exemple et qui sur son histogramme est très piquée en 0. Aussi nous avons la même conclusion que précédemment à la vue du Tableau 7 : les proportions sont significativement supérieures à celles attendues.

A nouveau, la sampled posterior predictive p-value a détecté des problèmes dans le modèle statistique nous faisant alors comprendre qu'il est à améliorer.

*Tableau 6 : résultats du test de Kolmogorov-Smirnov sur scénario 3*

| | Données non normalisées | Données normalisées |
|---|---|---|
| Moyenne | Statistique :0.04<br>p-valeur : 0.67 | Statistique :0.06<br>p-valeur : 0.18 |
| Variance | Statistique :0.28<br>p-valeur : < 2.2e-16 | Statistique :0.91<br>p-valeur : < 2.2e-16 |
| Kurtosis | Statistique :0.39<br>p-valeur : < 2.2e-16 | Statistique :0.73<br>p-valeur : < 2.2e-16 |
| Skewness | Statistique :0.39<br>p-valeur : < 2.2e-16 | Statistique :0.38<br>p-valeur : < 2.2e-16 |
| Variance/Moyenne | Statistique :0.31<br>p-valeur : < 2.2e-16 | Statistique :0.07<br>p-valeur : 0.07 |
| AIC(eff~site*esp) - AIC(eff~site+esp) | Statistique :0.43<br>p-valeur : < 2.2e-16 | Statistique :0.21<br>p-valeur : 3.52e-12 |

*Tableau 7 : Proportion de p-valeurs inférieures à 0.001, 0.01 et 0.05 pour le scénario 3*

| | Proportion p-valeur<0.001 | | Proportion p-valeur<0.01 | | Proportion p-valeur<0.05 | |
|---|---|---|---|---|---|---|
| Moyenne | Norm : 0.00333 | Non norm : 0 | Norm : 0.02 | Non norm : 0.014 | Norm : 0.08 | Non norm : 0.05666 |
| Variance | Norm : 0.312 | Non norm : 0.035 | Norm : 0.91 | Non norm : 0.15 | Norm : 0.9468 | Non norm : 0.246 |
| Kurtosis | Norm : 0.273 | Non norm : 0.1126 | Norm : 0.719 | Non norm : 0.3169 | Norm : 0.780 | Non norm : 0.397 |
| Skewness | Norm : 0.1028 | Non norm : 0.098 | Norm : 0.304 | Non norm : 0.30 | Norm : 0.421 | Non norm : 0.422 |
| Variance/Moyenne | Norm : 0.0035 | Non norm : 0.038 | Norm : 0.0106 | Non norm : 0.19 | Norm : 0.053 | Non norm : 0.313 |
| AIC(eff~site*esp)- AIC(eff~site+esp) | Norm : 0.042 | Non norm : 0.1338 | Norm : 0.1453 | Non norm : 0.36 | Norm : 0.223 | Non norm : 0.45 |

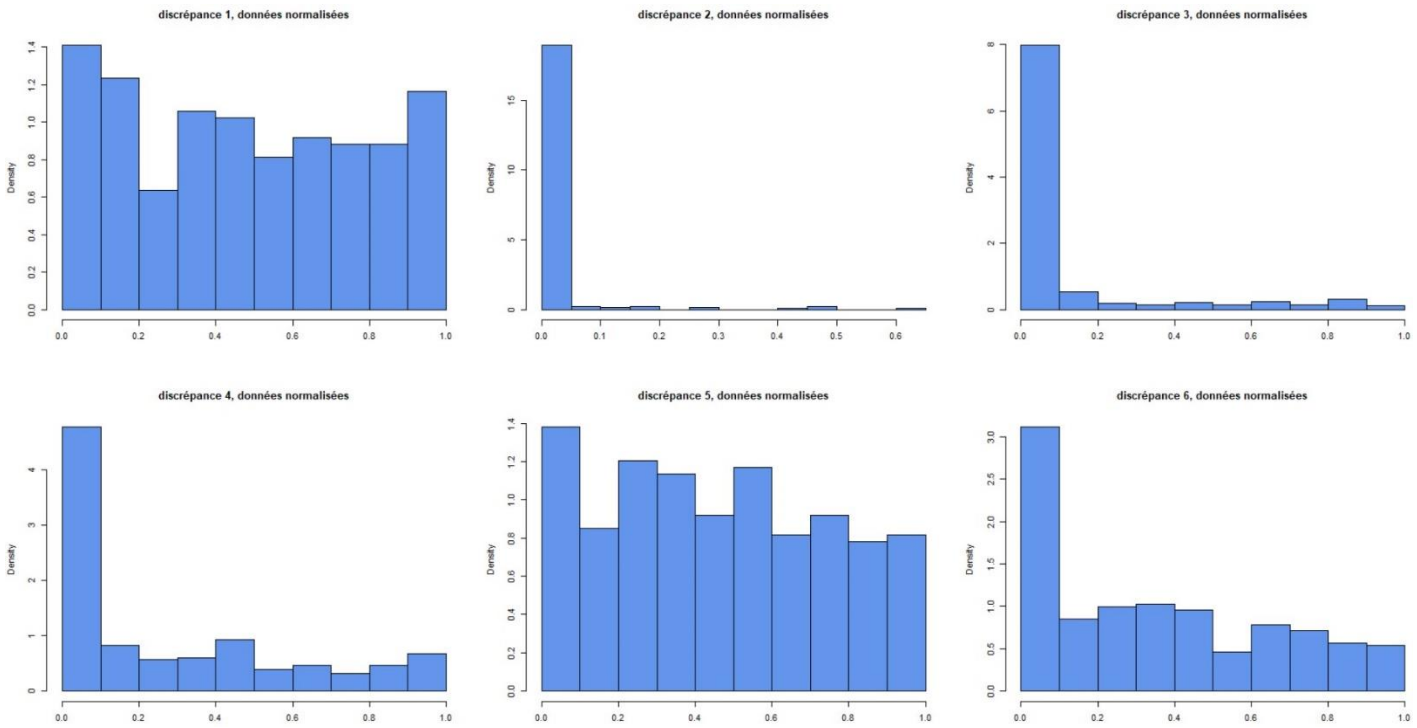


*Figure 7 : histogrammes du scénario 3 sur données non normalisées*

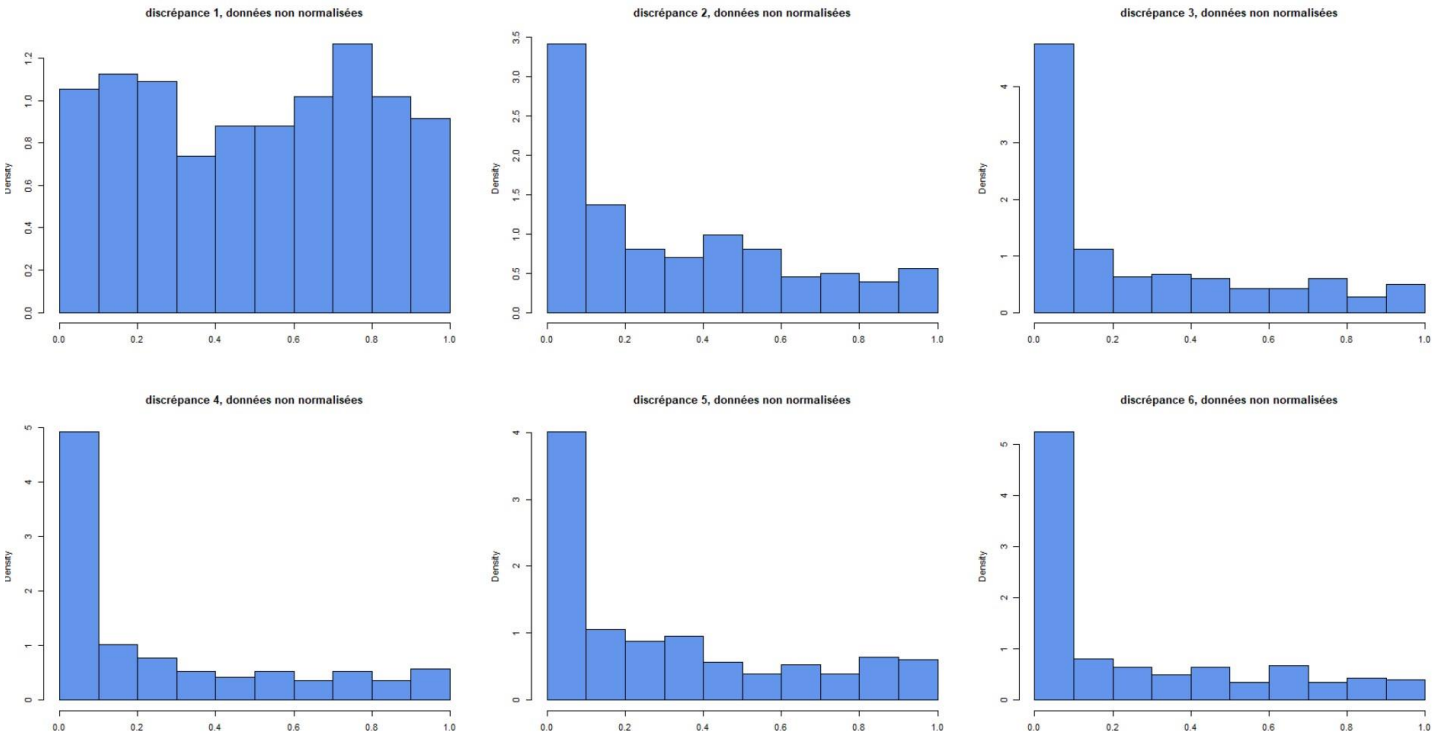


*Figure 8 : histogrammes du scénario 3 sur données normalisées*

## 3.4. Scénario 4

Pour ce quatrième scénario, chaque processus de 50 RData a mis 4 heures. On a encore le même diagnostic : toutes les fonctions de discrépances sont rejetées exceptée la moyenne sur données non normalisées. On remarque également dans la Figure 10 que la fonction de discrépance variance/moyenne est piquée en 1 et non en 0. D'ailleurs ce rapport a parfois des comportements qui peuvent *a priori* laisser penser à une loi normale ou une loi semblable (voir scénario 2 et 5).

*Tableau 8 : résultats du test de Kolmogorov-Smirnov sur scénario 4*

| | Données non normalisées | Données normalisées |
|---|---|---|
| Moyenne | Statistique :0.06<br>p-valeur : 0.27 | Statistique :0.68<br>p-valeur : < 2.2e-16 |
| Variance | Statistique :0.88<br>p-valeur : < 2.2e-16 | Statistique :0.98<br>p-valeur : < 2.2e-16 |
| Kurtosis | Statistique :0.68<br>p-valeur : < 2.2e-16 | Statistique :0.98<br>p-valeur : < 2.2e-16 |
| Skewness | Statistique :0.70<br>p-valeur : < 2.2e-16 | Statistique :0.68<br>p-valeur : < 2.2e-16 |
| Variance/Moyenne | Statistique :0.90<br>p-valeur : < 2.2e-16 | Statistique :0.52<br>p-valeur : < 2.2e-16 |
| AIC(eff~dataset+esp)-AIC(eff~1) | Statistique :0.93<br>p-valeur : < 2.2e-16 | Statistique :0.75<br>p-valeur : < 2.2e-16 |
| AIC(eff~dataset*esp)-AIC(eff~1) | Statistique :0.93<br>p-valeur : < 2.2e-16 | Statistique :0.87<br>p-valeur : < 2.2e-16 |

*Tableau 9 : Proportion de p-valeurs inférieures à 0.001, 0.01 et 0.05 pour le scénario 4*

| | Proportion p-valeur<0.001 | | Proportion p-valeur<0.01 | | Proportion p-valeur<0.05 | |
|---|---|---|---|---|---|---|
| Moyenne | Norm : 0.2 | Non norm : 0 | Norm : 0.5685 | Non norm : 0.000666 | Norm : 0.072 | Non norm : 0.0433 |
| Variance | Norm : 0.39 | Non norm : 0.29666 | Norm : 0.99 | Non norm : 0.8533 | Norm : 1 | Non norm : 0.9233 |
| Kurtosis | Norm : 0.46 | Non norm : 0.23 | Norm : 0.97 | Non norm : 0.6233 | Norm : 0.99 | Non norm : 0.72333 |
| Skewness | Norm : 0.2575 | Non norm : 0.24 | Norm : 0.62 | Non norm : 0.68666 | Norm : 0.729 | Non norm : 0.74666 |
| Variance/Moyenne | Norm : 0 | Non norm : 0.34 | Norm : 0 | Non norm : 0.89666 | Norm : 0.01 | Non norm : 0.93 |
| AIC(eff~dataset+esp)-AIC(eff~1) | Norm : 0.28 | Non norm : 0.39 | Norm : 0.69 | Non norm : 0.93 | Norm : 0.80 | Non norm : 0.963 |
| AIC(eff~dataset*esp)-AIC(eff~1) | Norm : 0.30 | Non norm : 0.37 | Norm : 0.85 | Non norm : 0.93 | Norm : 0.90 | Non norm : 0.963 |

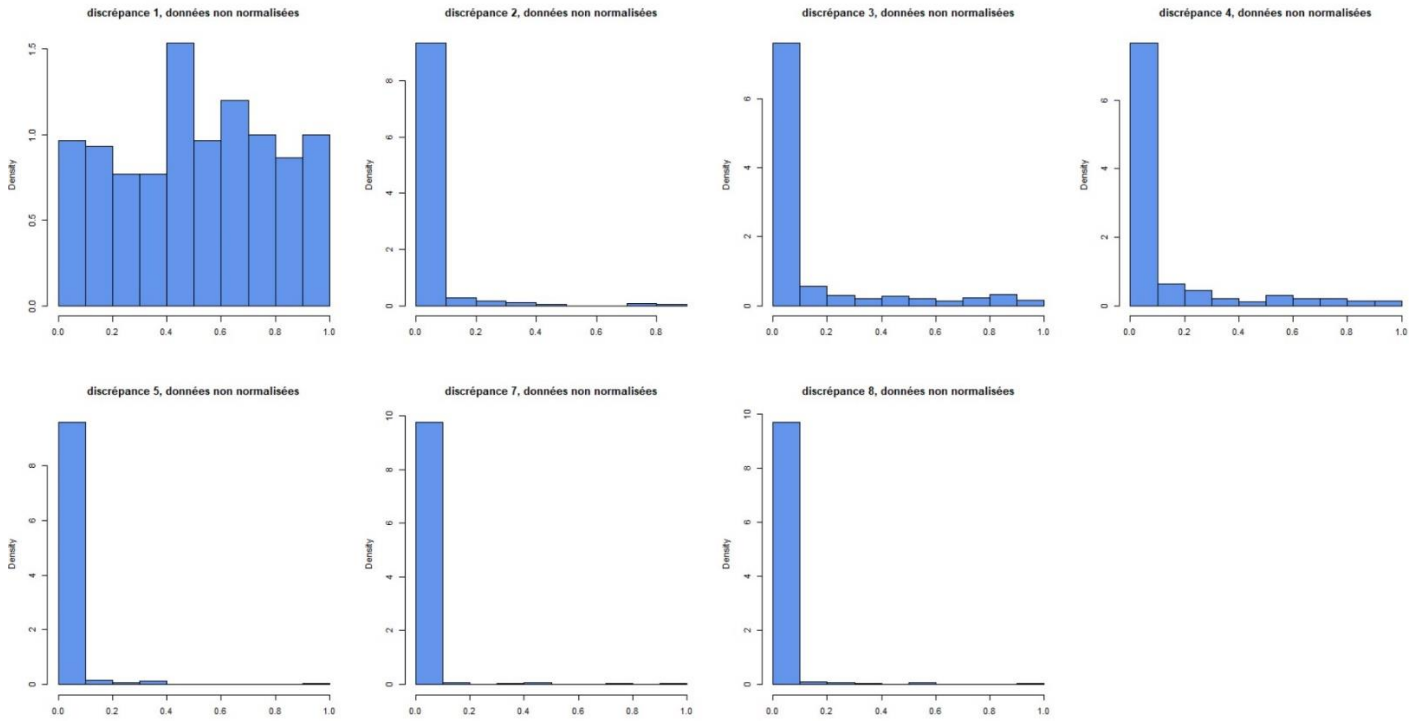


*Figure 9 : histogrammes du scénario 4 sur données non normalisées*

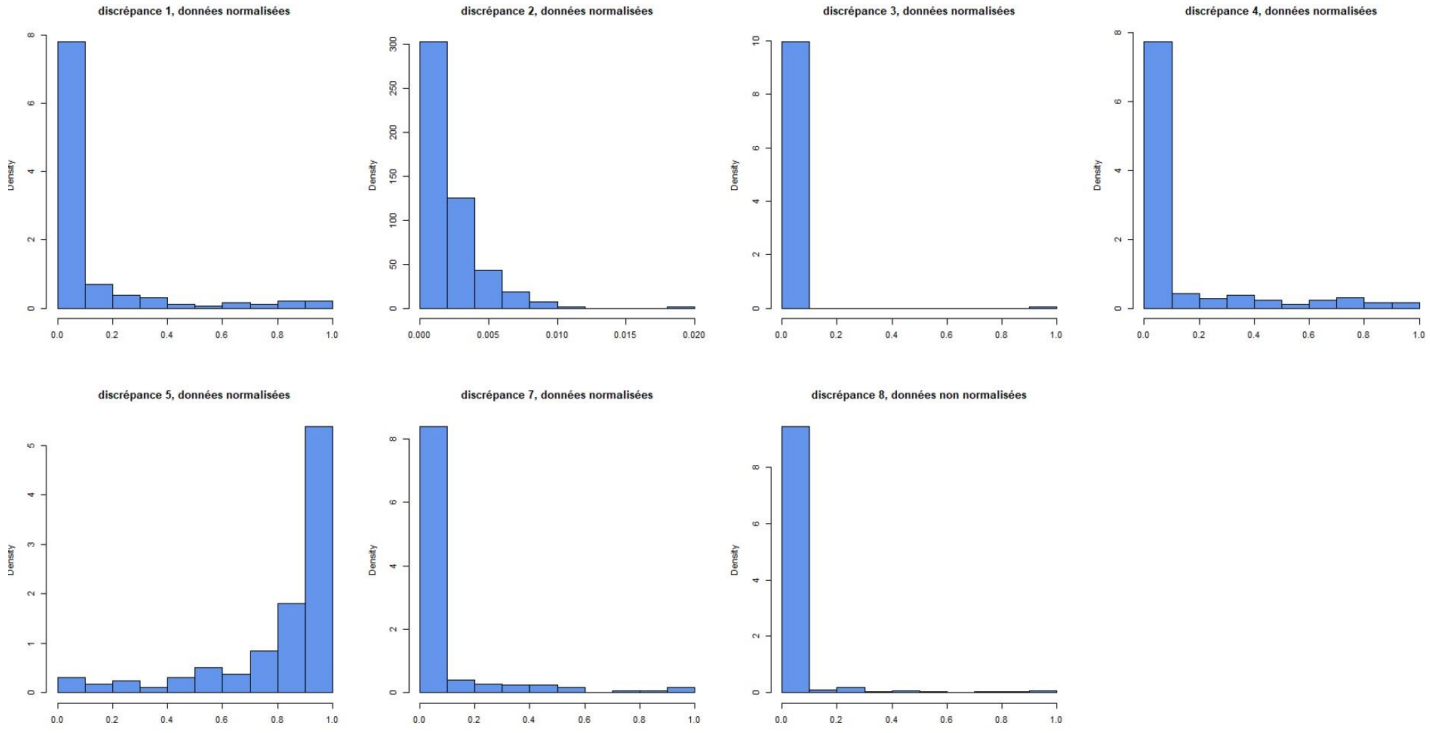


*Figure 10 : histogrammes du scénario 4 sur données normalisées*

## 3.5. Scénario 5

Pour ce cinquième et dernier scénario, chaque processus de 50 RData a mis 48 heures. On a encore le même diagnostic : toutes les fonctions de discrépances sont rejetées exceptée la moyenne sur données non normalisées. On remarque une nouvelle fois sur la Figure 12 le comportement en courbe de cloche qu'a la cinquième fonction de discrépance.

*Tableau 10 : résultats du test de Kolmogorov-Smirnov sur scénario 5*

| | Données non normalisées | Données normalisées |
|---|---|---|
| Moyenne | Statistique :0.05<br>p-valeur : 0.71 | Statistique :0.91<br>p-valeur : < 2.2e-16 |
| Variance | Statistique :0.93<br>p-valeur : < 2.2e-16 | Statistique :0.99<br>p-valeur : < 2.2e-16 |
| Kurtosis | Statistique :0.59<br>p-valeur : < 2.2e-16 | Statistique :0.98<br>p-valeur : < 2.2e-16 |
| Skewness | Statistique :0.59<br>p-valeur : < 2.2e-16 | Statistique :0.63<br>p-valeur : < 2.2e-16 |
| Variance/Moyenne | Statistique :0.94<br>p-valeur : < 2.2e-16 | Statistique :0.28<br>p-valeur : 5.32e-15 |
| AIC(eff~site)-AIC(eff~1) | Statistique :0.95<br>p-valeur : < 2.2e-16 | Statistique :0.71<br>p-valeur : < 2.2e-16 |
| AIC(eff~site*esp)-AIC(eff~1) | Statistique :0.97<br>p-valeur : < 2.2e-16 | Statistique :0.99<br>p-valeur : < 2.2e-16 |

*Tableau 11 : Proportion de p-valeurs inférieures à 0.001, 0.01 et 0.05 pour le scénario 5*

| | Proportion p-valeur<0.001 | | Proportion p-valeur<0.01 | | Proportion p-valeur<0.05 | |
|---|---|---|---|---|---|---|
| Moyenne | Norm : 0.29 | Non norm : 0 | Norm : 0.88 | Non norm : 0.0094 | Norm : 0.95 | Non norm : 0.05666 |
| Variance | Norm : 0.38 | Non norm : 0.32 | Norm : 0.99 | Non norm : 0.910 | Norm : 1 | Non norm : 0.952 |
| Kurtosis | Norm : 0.40 | Non norm : 0.22 | Norm : 0.98 | Non norm : 0.56 | Norm : 1 | Non norm : 0.636 |
| Skewness | Norm : 0.22 | Non norm : 0.15 | Norm : 0.58 | Non norm : 0.51 | Norm : 0.674 | Non norm : 0.627 |
| Variance/Moyenne | Norm : 0 | Non norm : 0.38 | Norm : 0 | Non norm : 0.938 | Norm : 0 | Non norm : 0.957 |
| AIC(eff~site)-AIC(eff~1) | Norm : 0.03 | Non norm : 0.35 | Norm : 0.2535 | Non norm : 0.948 | Norm : 0.569 | Non norm : 0.971 |
| AIC(eff~site*esp)-AIC(eff~1) | Norm : 0.36 | Non norm : 0.41 | Norm : 0.99 | Non norm : 0.976 | Norm : 1 | Non norm : 0.976 |

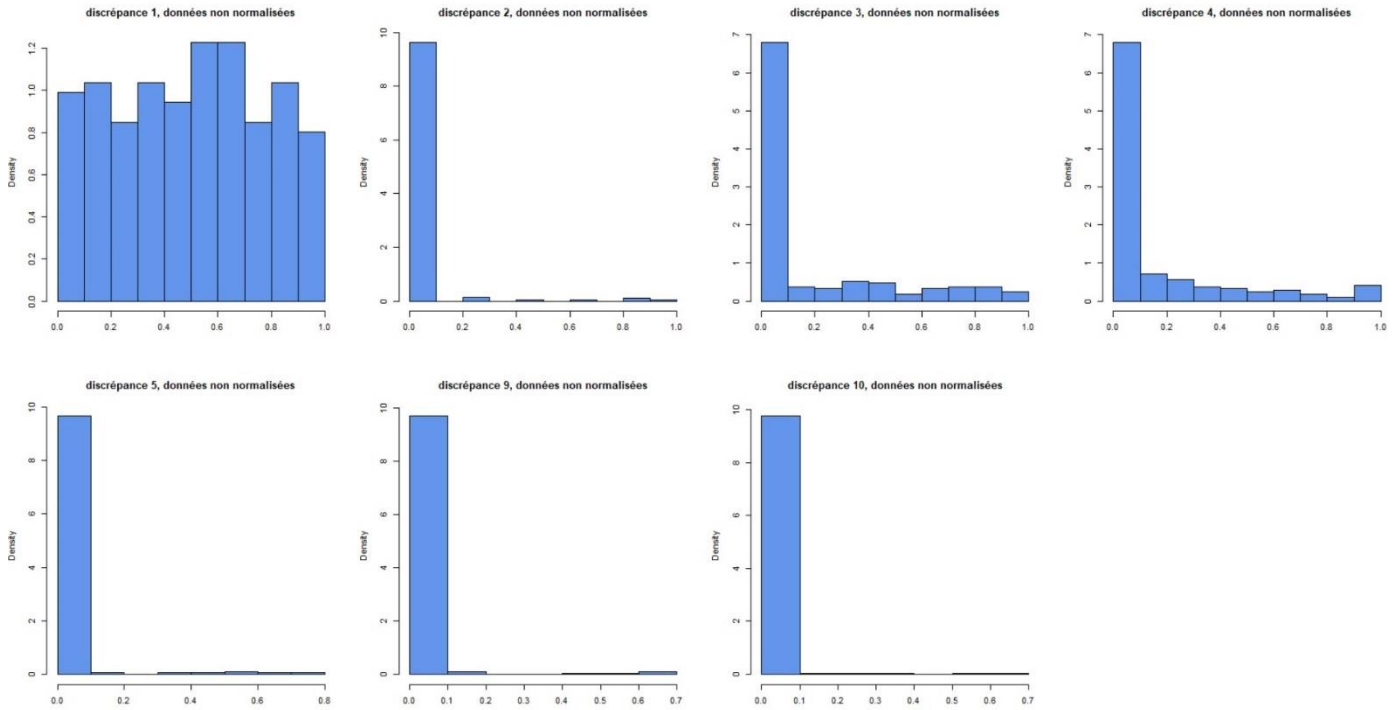


*Figure 11: histogrammes du scénario 5 sur données non normalisées*

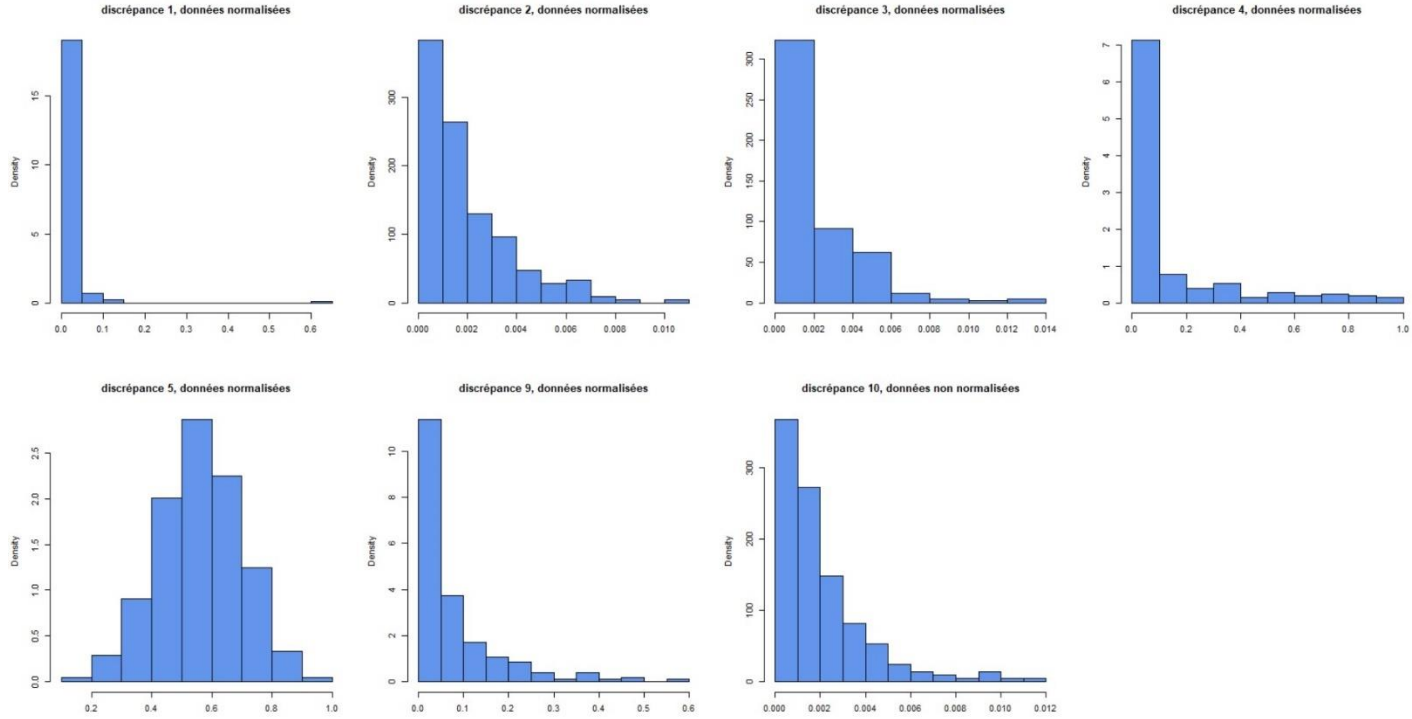


*Figure 12 : histogrammes du scénario 5 sur données normalisées*

# Discussion

On peut ainsi voir la puissance et l'intérêt de la sampled posterior predictive p-value dans un contexte de données non simulées. En effet celle-ci a détecté des problèmes dans tous les scénarii qui en présentent. Il faut aussi bien choisir les fonctions de discrépance. Toutefois nous avons été confrontés au problème de la « sur-puissance » de la sampled posterior predictive p-value. En effet comme nous l'avons dit dans « Boucle while et GOF » nous avons dû tirer au sort des observations de sorte à ne pas toutes les intégrer dans le calcul de la sppp. Le cas échéant, le trop grand nombre d'observations aurait mené au rejet même pour le scénario 1 car la sppp aurait détecté des problèmes infinitésimaux comme par exemple une moyenne valant 0.001 au lieu de 0. Ce problème aurait pu venir d'un mauvais choix de lois *a priori* par exemple et non d'une mauvaise spécification du modèle statistique entier. D'autres contextes cette puissance peut s'avérer utile mais ici nous ne devions pas fausser nos conclusions pour de si petites différences. C'est là où le bât blesse : la sppp peut s'avérer trop puissante à détecter des problèmes et en particulier arrive à en détecter certains qui écologiquement pourraient être négligeables. Il semblait également avoir des différences d'analyse entre le test de Kolmogorov-Smirnov et les histogrammes sur certains cas (scénario 3). En outre, nous avons eu des traces de divergences dans certains résultats. Par exemple sur les données normalisées certaines fonctions de discrépance rejettent sur le scénario 1. Il y a également le problème des données que l'on a dû tirer au sort, car en les prenant toutes on obtenait encore une fois des résultats plus extrêmes même pour le scénario 1. Malheureusement nous n'avons pas eu le temps de pleinement creuser ces divergences.

Aussi les limites de notre étude sont que nous avons travaillé uniquement sur des données simulées et n'avons pas appliqué tout le protocole sur des données réelles. Ce faisant, nous avons posé des hypothèses, certaines étant fortes écologiquement parlant, que l'on ne pourrait pas forcément vérifier dans un contexte plus appliqué. Par ailleurs, la question du nombre d'observations à sélectionner pour éviter le problème de « sur-puissance » de la sampled posterior predictive p-value se pose si l'on souhaite transposer à des données non simulées.

De plus, faute de temps, nous n'avons pas pu comparer d'une part la sppp avec uniquement les données protocolées et d'autre part la sppp comme nous l'avons faite, c'est-à-dire données protocolées et opportunistes. Ceci aurait permis de mieux rentrer dans le cadre du projet PASSIFOR2 évoqué en début de ce rapport. On aurait alors pu mieux apercevoir les apports positifs et négatifs d'un couplage de données protocolées et opportunistes. Il s'agit d'une piste à explorer par la suite, sachant que nos premiers essais sur la question semblaient indiquer que considérer les données standardisées seules montre des différences plus fortes sur scenario 1.

On peut également ajouter qu'il n'apparaît pas clairement si l'on peut détecter l'hypothèse violée à partir de la p-valeur la plus significative.

# Conclusion

Le rapport de stage touchant à sa fin, il est opportun de faire le bilan du travail réalisé d'une part mais aussi des apports de ce stage d'autre part.

Tout d'abord le but global était de mettre en place un outil pour détecter si les données opportunistes peuvent être adjointes aux données protocolées. (Coron et al, 2018.) ont proposé une méthode statistique pour ce faire sous certaines hypothèses. On a voulu d'abord vérifier que notre méthode – l'utilisation de la sampled posterior predictive p-value – ne détectait pas d'incohérence entre modèle statistique et données, dans le cas étudié par (Coron et al, 2018.). On a voulu ensuite s'assurer que notre méthode détectait des cas où certaines des hypothèses de génération des données (modèle probabiliste) sont changées.

Les résultats sont tout à fait encourageants et montrent que la sampled posterior predictive p-value permet de détecter des problèmes sur des données de simulation, malgré des traces de divergences qui sont encore à explorer. On pourrait alors envisager de transposer cette méthode sur des données réelles mais en prenant quelques précautions toutefois. Par exemple il faudrait se poser la question de la proportion de données à intégrer pour le calcul de la sampled posterior predictive p-value, relativement à la question de sur-puissance évoquée plus haut. Il faudrait également pouvoir vérifier les hypothèses faites par (Coron et al, 2018.) pour pouvoir utiliser leur modèle. Aussi dans un contexte d'utilisation pratique sur données réelles, on ne disposerait pas de 300 valeurs de GOF comme nous avons eu mais d'une seule. Or d'après nos calculs de proportions, sur certaines fonction de discrépances on aurait par exemple 25% de chances d'avoir une GOF qui détecte à raison un problème (cf Tableau 11). Il se pourrait qu'alors, dans un contexte de données réelles, on ne puisse pas détecter un problème existant.

Ensuite, abordons les apports de ce stage d'un point de vue plus personnel. Tout d'abord ce stage m'a initié à la recherche scientifique et plus particulièrement à la recherche en statistiques appliquées à l'écologie. Lire des articles, les recouper entre eux, se poser des questions, travailler en autonomie voire en autodidaxie, faire une bibliographie sont autant de choses que j'ai apprises et qui rythment le quotidien d'un chercheur. En outre j'ai pu utiliser des clusters de calcul, en particulier celui de Nogent-sur-vernisson. Bien entendu j'ai découvert et appris plein de nouvelles choses sur les statistiques bayésiennes et même les modèles statistiques utilisés en écologie comme les *Generalized Linear Mixed Models* (GLMM) ou encore la régression de Poisson. Par ailleurs plus je lisais et découvrais des choses plus je mesurais l'étendue des choses que j'ignorais. Cela semble être commun à plusieurs personnes en recherche et on peut utiliser la maxime de Socrate à ce sujet : « je ne sais qu'une chose, c'est que je ne sais rien ».

# Annexe n°1 : Etude de cas à propos de l'influence du nombre d'échantillons MCMC sur l'erreur de type I et le diagnostic de Geweke, et effet de l'autocorrélation sur le diagnostic de Geweke

Comme nous l'avons dit dans la partie principale, nous nous sommes questionnés sur le nombre d'échantillons MCMC à considérer. Pour mieux appréhender l'influence du nombre d'échantillons MCMC sur l'erreur de type I et le diagnostic de convergence de Geweke, nous allons construire trois modèles très simples et nous placer dans un contexte de simulation.

## Premier modèle : lois conjuguées

En ce qui concerne le premier modèle choisi, nous avons décidé d'utiliser des lois conjuguées pour avoir une loi *a posteriori* explicite plus facilement. Ainsi on considère $\theta \sim Beta(a = 2, b = 5)$ et $x_i|\theta \sim Ber(\theta)$. On a donc :

- $\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}1_{[0,1]}(\theta)$
- $f(x|\theta) \propto \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$
- $\pi(\theta|x) \propto \theta^{a-1+\sum x_i}(1-\theta)^{b+n-1-\sum x_i}1_{[0,1]}(\theta)$

Ainsi $\theta|x \sim Beta(a + \sum x_i, b + n - \sum x_i)$. On pose $n = 100$ et on considère 10000 réplicas, i.e, 10000 jeux de données *x* différents associés chacun à une valeur de $\theta$ tirée au sort dans $Beta(a = 2, b = 5)$. On considère quatre valeurs d'échantillons différentes : 100, 200, 500 et 1000. Ensuite, pour chaque réplica et pour ces quatre échantillons de la densité *a posteriori*, on procède à :

i. Calcul d'intervalles de crédibilité bayésiens (ICB) suivant différents quantiles et pour $\alpha \in \{10\%, 5\%, 1\%, 1\%, 1‰\}$.
ii. Comparaison du risque « observé » en se basant sur l'appartenance ou non de la vraie valeur du paramètre (connue car simulée) aux ICB au risque « théorique » attendu (valeurs de $\alpha$ ci-dessus).
iii. Calcul du diagnostic de convergence de Geweke.

Viennent enfin les analyses sur la base de 10000 sorties ci-dessus, l'erreur de type I associée au nombre de fois où le vrai paramètre est dans l'intervalle de crédibilité et pour le diagnostic de Geweke, comparaison de sa distribution, suivant le nombre d'échantillons, à la loi normale centrée réduite à l'aide d'estimateurs à noyaux et de tests de Kolmogorov-Smirnov et Jarque-Bera.

Les résultats sur l'erreur de type I sont condensés dans le Tableau 12. On constate alors que, si on souhaite utiliser une erreur de type I à 1%, 100 échantillons sont insuffisants car on triple presque le risque observé. Ensuite on voit sur la Figure 13 un exemple de comparaison entre la densité estimée par noyaux de Geweke et de la densité d'une loi normale. On y constate que la densité de Geweke, attendue d'une loi normale centrée réduite, a un mode plus aplati (platykurtique). C'est pourquoi on s'attarde ensuite sur le test de Jarque-Bera car il s'agit d'un test d'hypothèse nulle de normalité, et dont la statistique s'appuie sur l'asymétrie (skewness) et l'applatissement (kurtosis). On utilise également un test de Kolmogorov-Smirnov.
Les résultats obtenus avec les deux tests de normalité sont dans le Tableau 13 .

*Tableau 12 : erreur de type I observée en fonction de l'erreur de type I attendue et du nombre d'échantillons dans la densité postérieure pour le modèle à lois conjuguées Beta-Bernoulli*

| Nombre échantillons / Erreur attendue | 100 | 200 | 500 | 1000 |
|---|---|---|---|---|
| $\alpha = 10\%$ | 11.94% | 11.15% | 10.71% | 10.41% |
| $\alpha = 5\%$ | 6.87% | 6.16% | 5.65% | 5.34% |
| $\alpha = 1\%$ | 2.79% | 2.02% | 1.23% | 1.1% |
| $\alpha = 1‰$ | 2.15% | 9.3‰ | 2.87‰ | 2.1‰ |

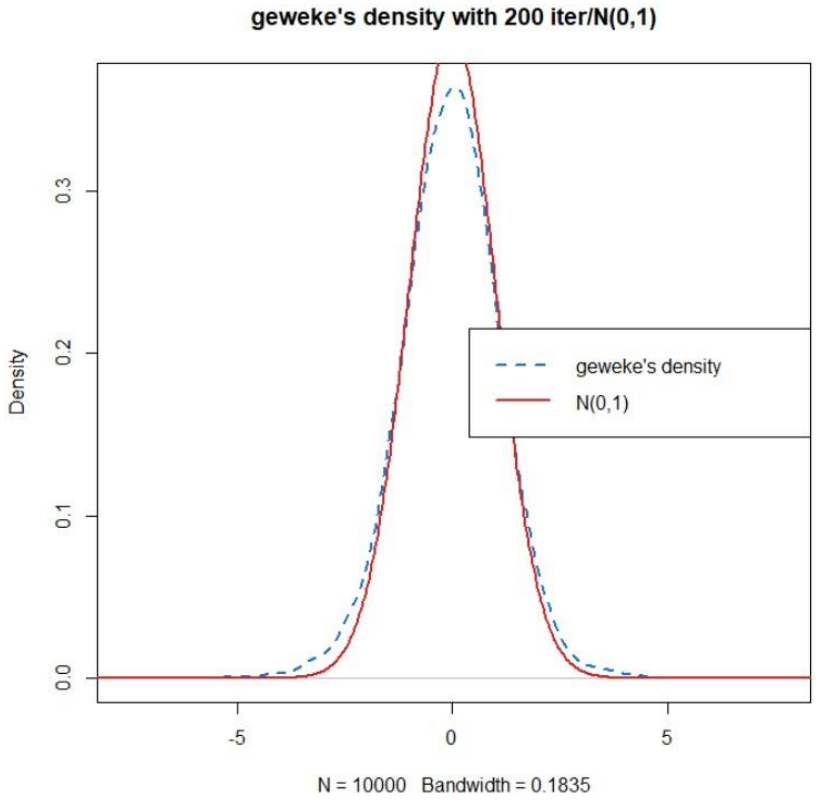


*Figure 13 : densité estimée par noyaux du diagnostic de convergence de Geweke avec 200 échantillons de densité postérieure et densité d'une loi normale centrée réduite*

*Tableau 13 : p-valeur suivant le test de normalité effectué et suivant le nombre d'échantillons de la densité postérieure pour le modèle à lois conjuguées Beta-Bernoulli*

| Nb echantillons / Test statistique | 100 | 200 | 500 | 1000 |
|---|---|---|---|---|
| Kolmogorov-Smirnov | $2.65 \times 10^{-13}$ | $3.39 \times 10^{-8}$ | $3.83 \times 10^{-3}$ | $8.3 \times 10^{-2}$ |
| Jarque-Bera | $< 2.2 \times 10^{-16}$ | $< 2.2 \times 10^{-16}$ | $< 2.2 \times 10^{-16}$ | $2.08 \times 10^{-5}$ |

On constate que le test de Kolmogorov-Smirnov conduit au non-rejet de la normalité au risque de 5% pour 1000 échantillons mais conduit au rejet au risque de 5% pour les autres cas. Le test de Jarque-Bera rejette toujours. Or nous utilisons le diagnostic de convergence de Geweke en le comparant à une loi normale centré réduite ! Les résultats de cette simulation annexe nous amèneraient à penser qu'alors pour avoir un diagnostic de convergence de Geweke au plus proche d'une loi normale, il faut beaucoup d'échantillons postérieurs.

## Deuxième modèle : modèle linéaire simple

Pour ce deuxième modèle nous allons utiliser le modèle linéaire bayésien simple : le modèle NIG (Normal, Inverse-Gamma). Plus précisément, nous simulons un jeu de données de taille 100 appelé x1 suivant une loi normale de moyenne 10 et de variance 9. Ce jeu de données est la variable explicative et la variable à expliquer est définie comme $y = 3 \times x_1 + \epsilon$ où $\epsilon \sim \mathcal{N}(0,1)$. Ainsi les paramètres sont $(\beta, \sigma^2) = (3,1^2)$. Nous allons prendre 4 échantillons de tailles différentes (100,200,500,1000) dans la loi *a posteriori* afin de voir l'influence du nombre d'échantillons sur l'erreur de type I. Nous prenons la loi $\mathcal{N}(0, \frac{1}{0.0001})$ comme loi *a priori* sur $\beta$ et la loi $IG(1, \frac{1}{1})$ comme loi *a priori* sur $\sigma^2$. Les résultats sont condensés dans le Tableau 14.

On remarque dans le Tableau 14 que le comportement de l'erreur de type I est sensiblement la même que pour le modèle précédent avec la loi Beta et la loi de Bernoulli.

*Tableau 14 : erreur de type I observée en fonction de l'erreur de type I attendue et du nombre d'échantillons dans la densité postérieure pour le modèle linéaire NIG*

| Nombre Echantillons / Erreur attendue | 100 | 200 | 500 | 1000 |
|---|---|---|---|---|
| $\alpha = 10\%$ | 11.05% | 10.83% | 10.17% | 10.22% |
| $\alpha = 5\%$ | 6.11% | 5.54% | 5.23% | 5.22% |
| $\alpha = 1\%$ | 2.04% | 1.31% | 1.25% | 1.13% |
| $\alpha = 1‰$ | 2.04% | 1.08% | 3.9‰ | 2.2‰ |

## Troisième modèle : AutoRegressif(1)

Dans cette partie nous allons nous attarder sur l'effet de l'autocorrélation des données sur le diagnostic de convergence de Geweke. Le but de cette simulation est de ressembler au contexte d'un MCMC avec des paramètres plus ou moins dépendants. Ainsi cela nous donnera une meilleur appréciation du comportement du diagnostic de convergence de Geweke. Pour cela nous allons faire deux scenarii de 10000 réplicas. Un premier scénario simule des données issues d'un modèle autorégressif d'ordre 1 avec une autocorrélation valant 0.2 et le deuxième scénario simule des données issues d'un modèle autorégressif d'ordre 1 avec une autocorrélation plus élevée valant 0.9. Pour chaque réplica :

i. Simuler 4 jeux de données de taille efficace variable (100, 500, 1000, 2000) qui sont issus du modèle AR(1) correspondant au scenario.
ii. Calculer pour ces 4 jeux de données leur nombre de valeurs efficaces avec rstan et en déduire une valeur de thin.
iii. Comparer la densité estimée par noyaux du diagnostic de convergence de Geweke sur les données issues de l'AR(1), la densité du diagnostic de convergence de Geweke sur les données « thinnées » et avec la loi de référence de ce dernier (N(0,1)).
iv. Calculer la proportion de valeurs de Geweke se trouvant dans l'intervalle [-1.96,1.96], correspondant à un intervalle de confiance de niveau 5% pour la loi normale centrée réduite.

On obtient par exemple avec un jeu de données de taille 500, pour l'autocorrélation valant 0.9 la Figure 14. On constate sur la Figure 14 le même comportement que l'on a vu précédemment, à savoir que le diagnostic de convergence de Geweke présente une densité plus aplatie (et donc présente des queues plus lourdes) que la loi normale centrée réduite. En revanche l'effet de l'autocorrélation est minime bien que les données « dépendantes » (i.e, non thinnées) semblent être plus pertinentes pour le score de Geweke. De plus, en prenant une taille d'échantillon plus grande, par exemple 1000, les densités de diagnostic de convergence de Geweke sont moins aplaties et donc plus proches de la loi normale centrée réduite (cf Figure 15).

En revanche, si on regarde les mêmes graphiques mais pour le scénario avec l'autocorrélation fixée à 0.2, on observe que le score de Geweke a la même densité pour les données thinnées et non thinnées (cf Figure 16 et Figure 17). Ce comportement n'est pas très surprenant car l'autocorrélation (fixée à 0.2) n'étant pas très forte, le thin est très faible et donc il y a presque le même nombre de valeurs entre les données thinnées et les données non thinnées. En outre la densité du diagnostic de convergence de Geweke sur données non thinnées avec 500 itérations est proche de la densité de Geweke sur données non thinnées avec 1000 itérations pour l'autocorrélation fixée à 0.9.

Pour finir on obtient le Tableau 15. Celui-ci corrobore les observations faites avec les figures à savoir qu'il est préférable de calculer le diagnostic de convergence de Geweke sur les données non thinnées.

Ainsi dans notre document principal, à la lumière de ces simulations, on a tout intérêt à calculer le diagnostic de convergence de Geweke sur les données non thinnées (ou du moins thinnées avec un thin fixé *a priori* sur le MCMC). En effet plus on a d'itérations et plus le score de Geweke se rapproche d'une loi normale centrée réduite. Néanmoins nous avons eu connaissance de (Cowles, Roberts, & Rosenthal, 1999) a posteriori. Cette référence préconise plutôt de diagnostiquer le burn-in sur un échantillon MCMC puis de faire tourner une nouvelle chaîne auquel on appliquera le burn-in estimé précédemment plutôt que de directement utiliser le burn-in sur la chaîne qui a servi à le diagnostiquer.

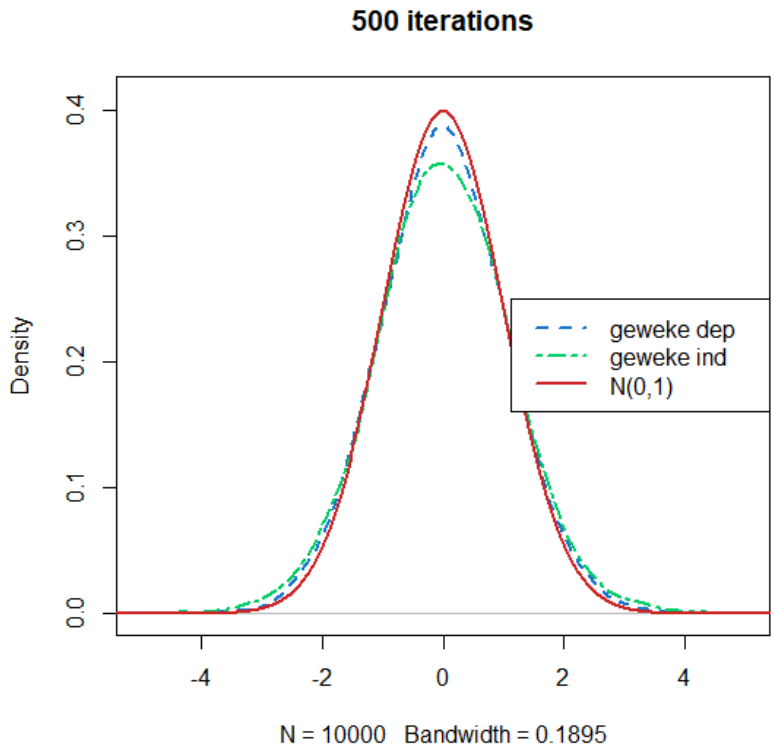


*Figure 14 : densité du diagnostic de convergence de Geweke pour les données autocorrélées (0.9), les données thinnées et la loi référence normale centrée réduite pour une taille d'échantillon de 500*

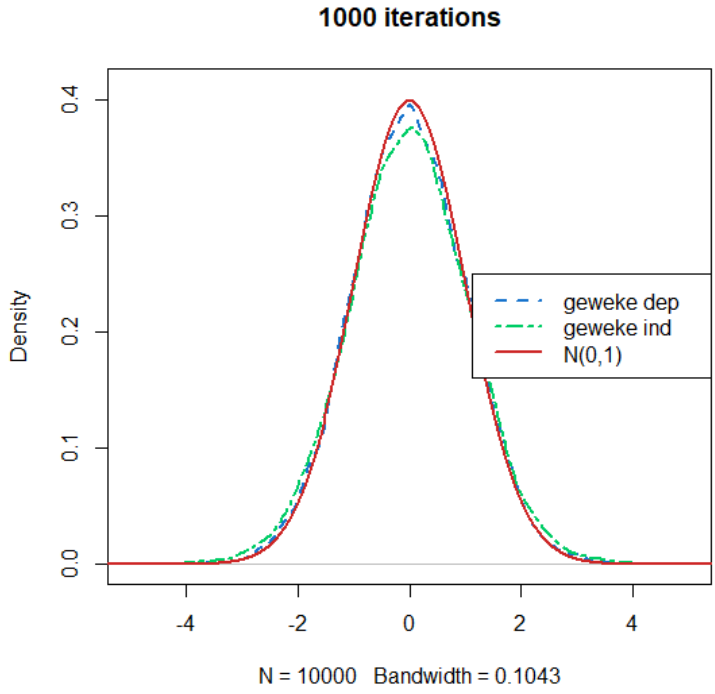


*Figure 15: densité du diagnostic de convergence de Geweke pour les données autocorrélées (0.9), les données thinnées et la loi référence normale centrée réduite pour une taille d'échantillon de 1000*

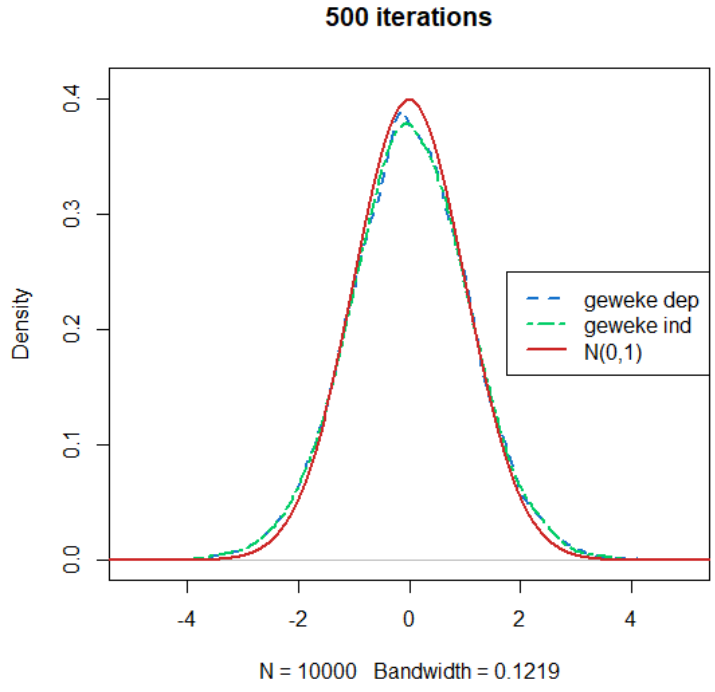


*Figure 16 : densité du diagnostic de convergence de Geweke pour les données autocorrélées (0.2), les données thinnées et la loi référence normale centrée réduite pour une taille d'échantillon de 500*

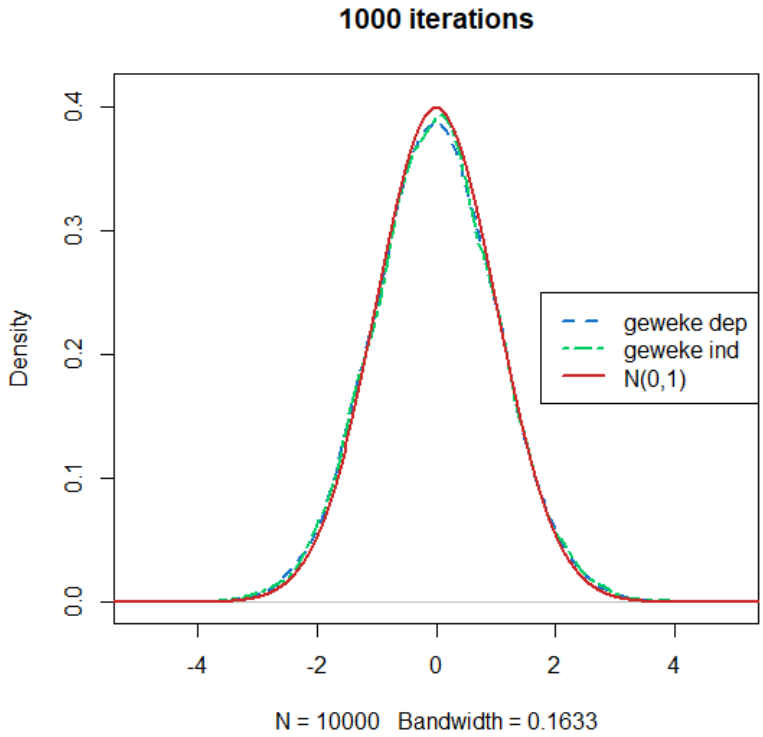


*Figure 17 : densité du diagnostic de convergence de Geweke pour les données autocorrélées (0.2), les données thinnées et la loi référence normale centrée réduite pour une taille d'échantillon de 1000*

*Tableau 15 : tableau récapitulatif des proportions observées pour chaque scénario suivant la valeur d'autocorrélation. La valeur de référence est 0.95*

| autocorrélation | 0.2 | | | | | 0.9 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Données thinnées | Valeurs efficaces | 100 | 200 | 500 | 1000 | Valeurs efficaces | 100 | 200 | 500 | 1000 |
| | proportion | 0.8816 | 0.8973 | 0.9289 | 0.939 | proportion | 0.8849 | 0.8976 | 0.9195 | 0.930 |
| Données non thinnées | Valeurs efficaces | 100 | 200 | 500 | 1000 | Valeurs efficaces | 100 | 200 | 500 | 1000 |
| | proportion | 0.8665 | 0.8905 | 0.9275 | 0.941 | proportion | 0.9053 | 0.9199 | 0.9377 | 0.946 |

# Références bibliographiques


Bayarri, M. J., & Berger, J. O. (2000). P-values for composite null models. *Journal of the American Statistical Association, 95*(452), 1127-1142. doi) Dec

Bayarri, M. J., & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science, 19*(1), 58-80.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association, 71*(356), 791-799.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A, 143*(4), 383-430.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician, 46*(3), 167-174.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician, 49*(4), 327-335.

Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician, 59*(2), 121-126.

Coron, C., Calenge, C., Giraud, C., & Julliard, R. (2018). Bayesian estimation of species relative abundances and habitat preferences using opportunistic data. *Environmental and Ecological Statistics, 25*(1), 71-93.

Cowles, M. K., Roberts, G. O., & Rosenthal, J. S. (1999). Possible biases induced by MCMC convergence diagnostics. *Journal of Statistical Computation and Simulation, 64*(1), 87-104.

Dennis, R. L. H., & Thomas, C. D. (2000). Bias in Butterfly Distribution Maps: The Influence of Hot Spots and Recorder's Home Range. *Journal of Insect Conservation, 4*(2), 73-77 %! Bias in Butter.

Evans, M. (2007). Comment: Bayesian checking of the second levels of hierarchical models. *Statistical Science, 22*(3), 344-348. doi:10.1214/07-sts235c

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian Data Analysis* (3rd ed.). Boca Raton: Chapman & Hall.

Geweke, J. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments* (Vol. 196): Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN.

Giraud, C., Calenge, C., Coron, C., & Julliard, R. (2015). Capitalizing on opportunistic data for monitoring relative abundances of species. *Biometrics, 72*(2), 649-658. doi:10.1111/biom.12431

Gosselin, F. (2011). A New Calibrated Bayesian Internal Goodness-of-Fit Method: Sampled Posterior p-values as Simple and General p-values that Allow Double Use of the Data. *PLoS ONE, 6*(3), e14770. doi:10.1371/journal.pone.0014770

Hjort, N. L., Dahl, F. A., & Hognadottir, G. (2006). Post-processing posterior predictive p values. *Journal of the American Statistical Association, 101*(475), 1157-1174. doi) Sep

Johnson, V. E. (2004). A Bayesian chi(2) test for goodness-of-fit. *Annals of Statistics, 32*(6), 2361-2384.

Johnson, V. E. (2007). Bayesian Model Assessment Using Pivotal Quantities. *Bayesian Analysis, 2*(4), 719-734.

Kass, R. E., Carlin, B. P., Gelman, A., & Neal, R. M. (1998). Markov chain Monte Carlo in practice: a roundtable discussion. *The American Statistician, 52*(2), 93-100.

Kéry, M., Royle, J. A., Schmid, H., Schaub, M., Volet, B., Häfliger, G., & Zbinden, N. (2010). Site-Occupancy Distribution Modeling to Correct Population-Trend Estimates Derived from Opportunistic Observations. *Conservation Biology, 24*(5), 1388-1397.

Kuussaari, M., Heliölä, J., Pöyry, J., & Saarinen, K. (2007). Contrasting trends of butterfly species preferring semi-natural grasslands, field margins and forest edges in northern Europe. *Journal of Insect Conservation, 11*(4), 351-366 %! Contrasting t.

Link, W. A., & Sauer, J. R. (1998). Estimating population change from count data: application to the north american breeding bird survey. *Ecological Applications, 8*(2), 258-268. doi:1998

O'Hagan, A. (2003). HSSS model criticism. In P. J. Green, N. L. Hjort, & S. T. Richardson (Eds.), *Highly Structured Stochastic Systems* (pp. 423-444): Oxford University Press.

Piccinato, L. (2000). Comments on Asymptotic distribution of P values in composite null models by J. M. Robins, A. van der Vaart and V. Ventura. *Journal of the American Statistical Association, 95*(452), 1166-1167. doi) Dec

Plummer, M. (2004). JAGS: Just another Gibbs sampler, URL : http://genome.jouy.inra.fr/applibugs/applibugs.07_11_08.plummer.pdf. In.

Plummer, M. (2019). rjags: Bayesian Graphical Models using MCMC. Rpackage version 4-10. https://CRAN.R-project.org/package=rjags. In.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC, R News, vol 6, 7-11. In.

Robins, J. M., van der Vaart, A., & Ventura, V. (2000). Asymptotic distribution of P values in composite null models. *Journal of the American Statistical Association, 95*(452), 1143-1156. doi) Dec

Schmeller, D. S., Henry, P. Y., Julliard, R., Gruber, B., Clobert, J., Dziock, F., . . . Henle, K. (2009). Advantages of volunteer-based biodiversity monitoring in Europe. *Conservation Biology, 23*(2), 307-316.

Snäll, T., Kindvall, O., Nilsson, J., & Pärt, T. (2011). Evaluating citizen-based presence data for bird monitoring. *Biological Conservation, 144*(2), 804-810.

Stan Development, T. (2020). RStan: the R interface to Stan. R package version 2.19.3. http://mc-stan.org/. In.

Szabo, J. K., Vesk, P. A., Baxter, P. W., & Possingham, H. P. (2010). Regional avian species declines estimated from volunteer-collected long-term data using List Length Analysis. *Ecological Applications, 20*(8), 2157-2169.

Team, R. D. C. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. In.

van Strien, A. J., van Swaay, C. A. M., & Termaat, T. (2013). Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. *Journal of Applied Ecology, 50*(6), 1450-1458.

Zhang, J. L. (2014). Comparative investigation of three Bayesian p values. *Computational Statistics and Data Analysis, 79*, 277-291.